ROYAL STATISTICAL SOCIETY

DATA | EVIDENCE | DECISIONS

BELFAST

2-5 SEPTEMBER 2019

RSS International Conference

ABSTRACTS BOOKLET

Headline sponsor

WILEY

Visit:

rss.org.uk/conference2019

#RSS2019Conf

Abstracts are ordered in date, time and session order for oral presentations followed by the poster presentations

1.1 Contributed - Medical: Data linkage

Tuesday 3 September 9am

Risk prediction models that use routinely collected electronic health data: generalisable and useful in heterogeneous settings?

Yan Li, Matthew Sperrin, Miguel Belmonte, Alexander Pate, Darren Ashcroft, Tjeerd Jeerd
Farr Institute for Health Informatics Research University of Manchester

Objective: To assess the extent of variability between practices on individual patients' risk of cardiovascular disease (CVD) that is not taken into account by the risk prediction model QRISK3.

Methods/Design: Longitudinal cohort study from 1st Jan 1998 to Jan 2015.Setting: 392 general practices (including 3.6 million patients) from the Clinical Practice Research Datalink (CPRD)Methods: Shared frailty model to incorporate QRISK3 predictors, practice variability and simulations to measure random variability.

Results: There was considerable variation in data recording between general practices. Practices on 5th percentile of missingness of Body mass index have 18.7% patients with missing values and 60.1% on the 95th percentile (for ethnicity, these were 19.6% and 93.9%, respectively). The crude incidence rates also varied considerably between practices (from 0.4 to 1.3 CVD events per 100 patient-years, respectively). The estimates of individual CVD risks with the random effect model were inconsistent with the estimated QRISK3 risk. For patients with a QRISK3 CVD risk of 10%, the 95% range of predicted risks were between 7.2% and 13.7% with the random effects model. Random variability only explained a small part of this. The random effects model was similar to QRISK3 for discrimination (C-statistic: 0.852 (95% CI: 0.850, 0.854)) and calibration (Brier score: 0.067 (95% CI: 0.067, 0.068)).

Conclusions: Risk prediction models that use routinely collected electronic health data can have limited generalisability and accuracy in predicting individual patient risks in heterogeneous settings. They need to be based on more robust evidence on causal risk factors.

1.1 Contributed - Medical: Data linkage

Tuesday 3 September 9am

Education and health outcomes of children treated for chronic conditions

Michael Fleming[1], James McLay[2], Danel Mackay[1], Jill Pell[1]
[1] *University of Glasgow,* [2] *University of Aberdeen*

Objectives: This retrospective cohort study linked Scotland-wide education data to national health data to explore associations between childhood chronic conditions and subsequent educational and health outcomes. Conditions studied were diabetes, asthma, epilepsy, attention deficit hyperactivity disorder (ADHD) and depression. We also explored neurodevelopmental multimorbidity (comorbid autism, learning disability, ADHD or depression). Educational outcomes were school absenteeism and exclusion, special educational need, academic attainment and subsequent unemployment. Health outcomes were hospital admissions and all-cause mortality.

Methods: Pupil census data and associated education records for all children attending primary and secondary schools in Scotland between 2009 and 2013 were linked to national prescribing data, acute and psychiatric hospital admissions, death records and retrospective maternity records enabling conditions to be studied whilst adjusting for sociodemographic and maternity factors and comorbid conditions. Conditions were ascertained from prescribing data and school records.

Results: All conditions were associated with increased school absenteeism, special educational need, and hospitalisation. All, excluding diabetes, were associated with poorer academic attainment and all, excluding ADHD were associated with increased mortality. ADHD and depression were associated with increased exclusion from school whilst epilepsy, ADHD and depression were associated with subsequent unemployment. Children experiencing neurodevelopmental multimorbidity had poorer outcomes across all educational domains. Depression was the biggest driver of absenteeism and ADHD was the biggest driver of exclusion.

Conclusion: In addition to poorer health outcomes, schoolchildren with these chronic conditions appear to experience significant educational disadvantage compared to their peers. Therefore interventions and further understanding of the intricate relationships between health and education among children with these conditions is required.

1.1 Contributed - Medical: Data linkage

Tuesday 3 September 9am

Evaluating the effects of an admission avoidance home visiting programme for frail, elderly patients in NHS Forth Valley

Maria Cristina Martin[1], Matt-Mouley Bouamrane[2], Kim Kavanagh[1], Paul Woolman[3]
[1] *University of Strathclyde,* [2] *University of Edinburgh,* [3] *NHS Forth Valley*

An increasing proportion of hospital inpatients are elderly, leading to the development of alternatives avoiding admission. Such interventions are often complex with several interconnecting parts. Their evaluation can prove challenging, with randomisation often being unfeasible. Routinely collected healthcare data provide an opportunity for evaluation in real-world settings, but carry several limitations. This research uses routinely collected retrospective data on elderly patients receiving admission avoidance home visits in NHS Forth Valley (FV). This submission presents the results of a matched cohort analysis and the difficulties of using real-world data for evaluation in this setting. Deterministic data linkage was used to compile an analytical dataset, linking several local and national datasets containing demographic and hospital activity data, for the entire elderly population in FV through Structured Query Language (SQL). The “Matching” R package was used to match intervention patients and controls. The outcome variable of whether an inpatient admission occurred or not in the year following intervention was evaluated using multivariate logistic regression.A cohort of 61,467 patients (566 of these intervention patients) were included. Intervention patients were found to be significantly different to general elderly patients on several variables, selected as matching variables for a control group (age, gender, Charlson comorbidity group, care home stay, living alone, and number of inpatient admissions experienced in the year prior to the intervention based on a dummy index date allocated to control patients) with 2:1 allocation. 477 intervention patients were successfully matched to 954 controls. Intervention patients were found to be more likely to experience an admission (OR 3.07, 95% CI 2.43-3.87) despite adjustment for residual differences. Despite matching, baseline hospital admissions for the two groups differed; hence, the selected matching variables may be insufficient for selecting an appropriate control group. Such issues in using routinely collected data for the purpose of evaluation will be discussed.

1.2 Contributed - Official and Public Policy: Quality and value in official statistics

Tuesday 3 September 9am

Hospital Standardised Mortality Ratio - Improving quality and safety of Scottish hospitals through the exploration and modelling of mortality

Robyn Munro, Lucinda Lawrie, David Caldwell
NHS NSS

OBJECTIVES:Information Services Division has produced quarterly Hospital Standardised Mortality Ratios (HSMR) for Scottish hospitals since December 2009. Data on hospital mortality have an important role to play in improving the quality of patient care. Over recent years there has been considerable debate about the utility and limitations of these measures. However, used wisely and as part of a wider suite of measures on quality and safety, these data can help drive improvements in patient care. In order to maximise the utility of HSMR, ISD have undertaken research to: Review the model methodology to ensure it continues to be appropriate and relevant;Interpret and disseminate key patterns shown in the data

METHODS:The HSMR is calculated by obtaining routinely collected death certificate data. These crude mortality data are adjusted to take account of some of the factors known to affect the underlying risk of death.The HSMR is calculated as: HSMR = Observed Deaths/ Predicted DeathsThe observed number of deaths is the total number of patients who died within 30-days of admission to hospital. The predicted number of deaths is calculated from a case-mix adjusted model.Using a three year dataset, logistic regression analyses were performed in order to test the relationship between explanatory variables and the outcome (whether the patient was alive or dead within 30 days). This involved fitting a model to data, evaluating fit, and estimating parameters that are later used in the prediction equation.

CONCLUSIONS:The existing HSMR methodology is robust and helps drive improvements in patient care; however to ensure comparisons which are made against the national average continue to be appropriate and relevant for each point in time the following recommendations were made:Regularly update the three year reference period used to model the relationship between explanatory variables and 30-day mortality.Update specific variables used for risk adjustment.

1.2 Contributed - Official and Public Policy: Quality and value in official statistics

Tuesday 3 September 9am

Reproducible Analytical Pipelines for Health and Social Care Publications

Jack Hannah, Anna Price
NHS National Services Scotland

Background: The Information Services Division (ISD) of the National Health Service Scotland produces approximately 200 health and social care publications annually. Most publications are produced using proprietary software such as SPSS, Business Objects and Microsoft Excel. Production of data and reports is time and labour intensive, involving extensive manual formatting and checking, and multiple movements of data between software. The Transforming Publishing Programme aims to modernise the publication process by creating Reproducible Analytical Pipelines (RAPs).

Methods: RAP combines the concept of reproducible research with data science best practices. It improves the quality, auditability and speed of publication production, as well as ensuring knowledge transfer in organisations with high turnover in staff. Our team focused on one publication as a proof of concept, developing an R package to produce the Quarterly Hospital Standardised Mortality Ratios (HSMR) publication. Git and GitHub were used for version controlling and facilitating collaborative working. We then developed a buddy system and toolkit to help other teams automate their reports, including an R style guide, project templates and RMarkdown publication templates. We also developed levels of code maturity and automation to help teams select an appropriate level to aim for, based on their skills and available development time.

Results and Conclusions: A standard RAP project includes each step of the publication production process: extraction from databases using SQL; data wrangling; documentation and unit testing of functions; and production of the final report using RMarkdown. The process of scaling RAP widely within the organisation has been facilitated by a buddy system whereby a RAP champion, who has previously worked on a RAP project, supports analysts from another team by providing expert advice, training and code reviews. This system has enabled the successful development of further RAPs at ISD and the opportunity for staff to develop code and new skills, with input and guidance from the Transforming Publishing team.

1.2 Contributed - Official and Public Policy: Quality and value in official statistics

Tuesday 3 September 9am

Dependent on Dover? Estimating and visualising the value of EU trade by UK port of entry for consumer goods

Jonathan Lewis
Civil Service

If a UK port were to suddenly become unusable, what is the most optimal route that a container of Radio equipment from Germany would take instead to import its products. What is the value of all Toys coming through Dover from Calais, and from which countries and EU ports did they originate. These are the questions that this project is aiming to answer, by developing a model that combines Economic and Statistical techniques to estimate the most optimal routes for products from over 20 EU directives, from every European Country, into the UK. We have then combined these probabilities with existing HMRC data to produce a Transport Mapper, which estimates how much trade is likely to be arriving into each UK port for each type of product.The model estimates which types of goods will prioritise time over cost and vice-versa in deciding the most optimal route, and a beta value will be used to adjust the sensitivity to using alternative routes on the final output. This evidence base does not yet exist, and could be very useful in understanding our trade flows with the EU into each UK port. It can for instance predict the next best alternative route to Dover in the expected scenario that there is disruption at Dover port.An interactive Transport Mapper dashboard has been developed which will allow policy makers to view those ports with the highest estimation of trade by product, along with the country of origin and which EU port it arrived from. This Dashboard will be live at the presentation to showcase the work.

1.3 Contributed - Environmental / Spatial Statistics:

Tuesday 3 September 9am

Combining citizen science and survey data in a log-Gaussian Cox process framework to estimate the monthly space-use of Southern Resident Killer Whales

Joe Watson, Marie Auger-Methe
University of British Columbia

Species distribution models (SDMs) are useful tools to help ecologists quantify species-environment relationships, and they are being increasingly used to help determine the impacts of future climate and habitat changes on species. Estimating SDMs can be tricky from a statistical point of view since the effects of spatial and temporal autocorrelations, land cover and environmental covariates and detectability functions all need to be considered and inherently modeled. Furthemore, such models often assume that data have been collected from well-designed surveys and/or studies. In practice, data are often of the form of presence-only sightings collected from 'citizen scientists' and/or industry, and their 'search effort' can be difficult to quantify. Furthermore, search effort from such sources is often concentrated in areas in which the expected count of the species under study is high, and/or where population density is high. Ignoring the search effort can lead to severely biased estimates of the species distribution.We look at data collected on Southern Resident Killer Whales (SRKWs), an ecotype with designated 'species at risk' status found off the coast of Vancouver Island. Data from from a variety of 'citizen science' sources and government surveys are considered. We present a method to combine the different data sources and estimate the monthly SRKW space-use in a statistically-rigorous manner using spatio-temporal log-Gaussian Cox processes within the R-INLA and inlabru packages. Improved (effort-corrected) estimates of the SRKW distribution will hopefully help ecologists and policy-makers safeguard the future of the SRKW.

1.3 Contributed - Environmental / Spatial Statistics:

Tuesday 3 September 9am

Detecting and Communicating changes in Waste Water Treatment Plant performance in Ireland

Jason Larkin
Environmental Protection Agency Ireland

A small team at the Environmental Protection Agency regulates Waste Water Treatment Plants (WWTPs) in Ireland. We collect hundreds of thousands of environmental monitoring data points each year relating to the performance of these plants. The Analytics team is striving to turn this hard monitoring data into information about how each plant is doing, and how it is impacting the wider environment.This presentation will show how we scaled from a small investigation into the performance of a single WWTP, to generating a set of metrics that measure the performance of each of ~500 WWTPs. These metrics use simple statistical tools like t-tests and sign-tests to summarise the monitoring data.It will also show how we communicate our metrics in an r-shiny dashboard.

1.3 Contributed - Environmental / Spatial Statistics:

Tuesday 3 September 9am

Understanding model fit for simulating species dispersal using alternative cost metrics

Laura Merritt[1], Justin Travis[2], Steven White[3], Tom Oliver[4], Rob Salguero-Gomez[5], James Bullock[3]

[1] University of Reading/Centre for Ecology and Hydrology, [2] University of Aberdeen, [3] Centre for Ecology and Hydrology, [4] University of Reading, [5] University of Oxford

Calculating measures of dispersal such as 'wave-speed' through integrodifference equations relies on the fitting of probability density functions called dispersal kernels. However, these estimates are particularly sensitive to the underestimation of rare long-distance dispersal events. Visual inspection of graphs showing fitted dispersal kernels compared to the original count data seemed to suggest consistent underestimation of long distance dispersal events. The objective of this study was to statistically quantify the difference in fit between these rare events, and shorter dispersal distances. Due to the magnitude of difference in counts between the first and last measured points, a novel cost metric was developed based on proportional differences between observed and predicted values. We conclude that rare long-distance dispersal events are being underestimated in dispersal kernels, while there is little error in shorter distances. We therefore suggest methods for fitting that can limit this underestimation and provide more accurate predictions of species dispersal speeds.

1.4 Contributed - Social Statistics: Neighbourhoods

Tuesday 3 September 9am

Life at the Frontier: Conceptualising the Causes and Consequences of Ambient Social Frontier Propensity

Gwilym Pryce
University of Sheffield

Quantitative research on segregation has focussed primarily on measuring the overall degree of separation within and across neighbourhoods, rather than on understanding the nature of the spatial transition from one neighbourhood to another. Yet, it is the nature of the residential interface between separate communities that might be of greatest importance. This paper aims to extend the theory of “social frontiers” and their impacts. Social frontiers are defined sharp spatial divisions in the residential make-up of adjacent communities (Dean et al. 2018), as opposed to more gradual blending of groups. Such frontiers represent invisible “cliff edges” in the socio-economic geography of neighbourhoods. We conceptualise social frontier formation through the lens of residential sorting approaches and reflect on their impacts on human wellbeing through a variety of complementary mechanisms. A series of hypotheses are proposed regarding the potential impact of proximity to social frontiers on educational attainment and life outcomes. We argue that, to properly account for such effects, we need to consider a number of important overlooked aspects of social frontiers, including the ambient propensity toward social frontiers in the surrounding areas and in wider society. This leads us to propose the concept of “Ambient Frontier Propensity” and to suggest a Bayesian statistical framework for estimating it.

1.4 Contributed - Social Statistics: Neighbourhoods

Tuesday 3 September 9am

Career Satisfaction, Work Resources and Health of Employees and of Their Children: Evidence from 1,883 Chinese Dual-Earner and Only-Child Households

Chunyi Chen, Mengjie Xu, Liang Guo
Shandong University

Prior studies have paid considerable attention to the relationship between work and family and there is a consensus that work can affect employee's mental and physical health. However, little attention has been devoted to analyzing how parents' work affects their children's health. This study tends to fill in this gap. Our samples include 1,883 dual-earner and only-child households in China. We follow the path analysis approach to estimate a series of regression models. Our study reveals that perceiving to gain valuable resources at work significantly promotes not only the health of working parents (0.077, $p<0.01$) but also the health status of their children (0.058, $p<0.05$). We also find that career satisfaction was positively related to better parental health (0.094, $p<0.01$), but it has a long-arm, negative effect on children's health (-0.070, $p<0.01$). These effects are partially mediated by parental health. Our study contributes to the work-family literature in the following ways. Firstly, we provide both theoretical and empirical justifications for the fact that parental career satisfaction promotes parental health but jeopardizes the health of their children. Secondly, we find that work resources promote the health of both parents and children. And finally, our sample size is the largest of its kind, representing both urban and rural households working in multiple industries. The findings are robust and may be generalized across different social-economic status.

1.4 Contributed - Social Statistics: Neighbourhoods

Tuesday 3 September 9am

Neighbourhood change in Britain, 1971-2011

Chris Lloyd[1], Gemma Catney[1], Paul Norman[2], Nick Bearman[3]
[1] *Queen's University Belfast,* [2] *University of Leeds,* [3] *University College London*

The paper brings new perspectives on major demographic and socioeconomic trends witnessed in Britain over the last 40 years – counterurbanisation and return to cities, a north-south divide in population growth, the decline of social housing, increases in overcrowding, the decline of manufacturing and mining, and the growth of the service sector. The analysis makes use of 1km gridded counts derived using Census data for 1971, 1981, 1991, 2001 and 2011 for Britain. The data were produced as a part of the ESRC-funded 'PopChange' project. The project entailed transfer of counts from source geographies (enumeration districts or output areas) to 1km grid cells to produce geographically-consistent data. The analyses presented are part of a large programme of work which has explored changes in population densities, deprivation, health, housing tenure, and housing spaces. The paper outlines the approaches used to construct the dataset before moving on to assess major geographical trends and local complexities in terms of age, country of birth, qualifications, employment, industry, housing tenure, and overcrowding. The analyses reveal marked changes in the geography of age profiles between 1971 and 2011, with a decline in the diversity of age groups in many neighbourhoods, including much of London. Other major temporal patterns include increased overcrowding in the south east of England, and considerable declines in manufacturing and mining - illustrated in far more geographical detail than in previous analyses. Common changes (variables which increase or decrease together) are explored using multidimensional scaling of correlation matrices for percentage point change variables. A final stage presents a classification schema developed using -k-means, allowing exploration of how neighbourhoods change in multiple ways over the 40-year period covered. The paper concludes by considering how the analyses can be extended beyond 2011, and it also highlights ways in which analyses of neighbourhood change can be vital in assessing the needs of local communities.

1.5 Contributed - Methods and Theory: Causal Inference

Tuesday 3 September 9am

Bayesian nonparametric estimation in longitudinal mediation: A Baron-Kenny based estimator for cross-lagged models

Andrej Srakar
Institute for Economic Research (IER), Ljubljana and Faculty of Economics, University of Ljubljana

There has been a significantly raised interest in causal inference in both statistics and econometrics in recent decades. Mediation analysis has its roots in the literature of structural equation models, both in the context of linear (MacKinnon, 2008; Glynn, 2012) as well as nonlinear (Rubin, 1974; Holland, 1988; Halpern, 1998; Pearl, 2001) models. Limitations of cross-sectional models to analyze mediation effects have been pointed in the literature and can be overcome by longitudinal modelling (Gollob and Reichardt, 1987; Selig and Preacher, 2009; Bernal Turnes and Ernst, 2016) which is the particular reason for studying longitudinal mediation, apart from it being a highly uncommon and underresearched methodology. Also, existing models for longitudinal mediation are estimated under strong parametric assumptions which imposes statistical problems (Bernal Turnes and Ernst, 2016).We derive a nonparametric Bayesian estimator for cross-lagged longitudinal mediation models (based on »classical« Baron-Kenny approach to mediation, see Baron and Kenny, 1986). As longitudinal mediation for cross-lagged models demands a dynamic panel modelling having lagged variables in the nonparametric part, we follow Su and Lu (2013) using iterative local kernel-based approach with sieves as initial estimators. To map to a Bayesian nonparametric "space" we use objective Dirichlet prior, recommended in the literature (Alvares, Armero and Forte, 2018) while also exploring in Bayesian nonparametric literature more commonly used Dirichlet Process Mixtures. We show the constructed estimator attains optimal information rate (Alaa and van der Schaar, 2018). We explore the properties of the approach in a simulation study, comparing the performance to the existing (parametric) estimators for cross-sectional and longitudinal mediation. In a short application, we study the mediated effects of cultural policy funding on the performance of nascent cultural firms using Amadeus firm-level data in the period 2007-2016.The article is the first exploration of Bayesian nonparametric approach in longitudinal mediation literature providing this field with important statistical considerations for its future development.

1.5 Contributed - Methods and Theory: Causal Inference

Tuesday 3 September 9am

Interpreting estimates of mediated effects from studies with attrition: an example from a study of maternal depression and child neurodevelopment.

Nicola Fitz-Simon[1], Alina Rodriguez[2]
[1] *National University of Ireland Galway,* [2] *University of Lincoln and Imperial College London*

Depression during pregnancy is common, affecting about 15% of pregnant women. Depression has been linked to compromised foetal development, including the foetal brain, via biological mechanisms, and is related to behavioural, emotional, and cognitive problems, i.e. neurodevelopmental deficits, in childhood.We analysed a longitudinal study of births in Sweden that recruited n=3,046 pregnant women at antenatal clinics in 1999 and 2000. We found a positive association (adjusted for baseline covariates) between maternal depression during pregnancy and child behavioural problems at age 5. By considering hypothetical interventions on alternative mediating pathways via postnatal maternal mental health and infant temperament, we estimated interventional direct and indirect effects, using Monte Carlo simulation. We found the adjusted association between pregnancy depression and child behaviour at age 5 was largely explained by postnatal maternal depression, thus suggesting little evidence of a foetal effect. However, this study was affected by attrition, with 5 year follow-up data available for only 55% of the cohort. Maternal depression, both during pregnancy and the postnatal period, was associated with 5-year attrition. Other variables associated with attrition were markers of socioeconomic position, parity, and stressful life events, which are potential confounders of the mediator-outcome associations. To investigate the impact of attrition on the mediated effect estimate, we did a series of sensitivity analyses. We found that an unmeasured confounder of plausible magnitude, that also predicts attrition, can give rise to a biased estimate of the mediated effect. Moreover, this bias is increased with increasing attrition. Mediated effect estimates in this study with high attrition must be interpreted with caution.

1.6 Contributed - Medical: Risk Factors

Tuesday 3 September 9am

Evaluation of two-part models for semi-continuous patient reported outcome measures: an application to a clinical trial of lower back pain.

James Griffin[1], Ranjit Lall[2], Jane Hutton[1]
[1] Department of Statistics, University of Warwick, [2] Warwick Clinical Trials Unit, University of Warwick

Patient reported outcome measures (PROMs) are extensively used to assess patients' perspective on impairment and pain in clinical studies and medical practice; and are often the primary outcome in randomised controlled trials of new treatments. PROMs sometimes have a point mass at one end of the scale, and skewed values over the remaining values and lead to heteroscedastic errors and non-constant variance. Hence conventional linear regression approaches; even with mathematical transformations applied, are invalid. In econometrics such distributions are termed "semi-continuous" as they are neither exclusively discrete nor continuous.We consider a two-part model approach to address these problems in semi-continuous PROMs. Econometricians have used such models for decades but there have been very limited application in medical statistics and clinical trials. Two-part models require treating the observed PROM values as arising from two distinct stochastic processes; a binary outcome taking value 0 if the score was zeroes and 1 if the score was non zero (the first part); and another governing the non-zero values conditioning on a non-zero response (the second part).We compare specifications of two-part models to predict treatment effects using the Roland Morris Disability Questionnaire in the Back Skills Training Study. Models studied included logistic and probit models for the first part, with transformed regression and generalised linear models for the second part.We present an approach to assess model fit, prediction errors and the validity of model assumptions which extends traditional residual checking approaches for conventional one part regression models. We also discuss how to derive and evaluate treatment effects in the complex setting of a two-part model and demonstrate the inadequacies of single number summaries.Our results demonstrate that two-part models are an interpretable and practical family of models and with careful consideration they should be an attractive option to analyse semi-continuous PROMs.

1.6 Contributed - Medical: Risk Factors

Tuesday 3 September 9am

Quantifying effects of some socio-demographic risk factors on Lyme disease incidence in Scotland.

Jude Eze[1], Roger Evans[2], Harriet Auty[1], Rita Ribeiro[1], Sally Mavin[2], Roger Humphry[1], Gunn George[1]
[1] SRUC, [2] NHS Highland

Problem: Lyme disease (LD) is a bacterial infection transmitted to humans through a bite by Ixodes ricinus tick infected with Borrelia burdgeferi. It is the most significant vector-borne disease in Europe and America with increasing number of cases reported annually. If left untreated, LD could lead to debilitating disease and serious morbidity with associated human and health costs. Variations in the incidence of LD over time and space may depend, among other factors, on the socioeconomic and demographic characteristics of the population. These characteristics may be strongly associated with behaviors and activities which expose individuals to risk of infection. LD prevalence studies have focused mainly on environmental and climatic factors. Few studies have indicated differences in prevalence rates among age groups and sex. This study aims at using advanced statistical models to obtain adjusted estimate of effects of patient's socioeconomic and demographic characteristics after accounting for correlation, trend and seasonal effects.

Methodology: Laboratory confirmed cases of LD between 2006 and 2017, with patient demographics, were supplied by National Lyme Borreliosis Testing Laboratory (NLBTL), Scotland. Using Scottish Index of Multiple Deprivation (SIMD) quintiles (1 = most deprived to 5 = least deprived) as proxy for socioeconomic status, we developed a multivariable logistic mixed effect model that accounted for within health board correlation.

Findings: Incidence of LD was strongly positively associated with patient's age and socioeconomic status. Increasing age by one year increased the odds of infection by 15% in females and 9% in males. Odds of infection among the most affluent were 1.33 times that of the most deprived. Being a male increased the odds of infection by 1.76 times relative to being female.

Conclusion: Being older, or male, or on the higher rungs of socioeconomic status, means greater odds of LD infection. This highlights the need to target individuals with these characteristics in any campaign to reduce LD infection.

1.6 Contributed - Medical: Risk Factors

Tuesday 3 September 9am

Incorporating misclassification error from finite mixture models into generalised linear models: an illustration from serological survey of RSV in England

Ania Zylbersztejn[1], Lucy Pembrey[2], Harvey Goldstein[1], Guy Berbers[3], Rutger Schepp[3], Fiona van der Klis[3], Charles Sande[4], Dan Mason[5], Rosalind Smyth[1], Pia Hardelid[1]
[1] UCL Great Ormond Street Institute of Child Health, London, UK, [2] London School of Hygiene and Tropical Medicine, London, UK, [3] National Institute of Public Health and the Environment (RIVM), Bilthoven, The Netherlands, [4] Kemri-Wellcome Trust Research Programme, Kilifi, Kenya, [5] Bradford Institute for Health Research, Bradford, UK,

Background: Bronchiolitis due to respiratory syncytial virus (RSV) is a leading cause of hospitalisations in infants. Community-based serological surveys of RSV can inform vaccine strategies but require costly, time-consuming and invasive repeated blood sampling. Instead, we estimate burden of RSV through secondary analysis of stored blood samples, using a two-stage modelling strategy based on finite mixture models (FMM).

Methods: We used stored blood samples, tested for RSV IgG antibodies, linked to questionnaires, primary and secondary care records from the Born in Bradford cohort study, collected at 1 and 2 years of age. We fitted FMM to antibody data and classified children as infected according to their posterior probability of infection. We estimated risk ratios for infection according to age, ethnicity, date of blood sample and indicators of population mixing (e.g.: childcare attendance) using Poisson regression with robust error variances. To account for possible misclassification of infection status, we simulated infection status according to the posterior probability of infection from the FMM and re-estimated risk ratios 50 times. We pooled the estimates of risk ratios using Rubin's rules.

Results: The study included 477 children. 249 (52%) children had serological evidence of RSV infection in infancy; 94 of whom (38%) had been in contact with healthcare during RSV season. Date of sampling, having older siblings, Pakistani ethnicity, and attending formal childcare were predictive of RSV infection in infancy. The infection status was well defined – 462 (97%) children had posterior probability of infection <10% or >90%. On average, 248 children (standard deviation=2) were indicated as infected across simulated datasets. Pooled risk ratios were comparable with the main analyses.

Conclusions: This is the first community-based serological survey of RSV in England. Secondary use of stored serological data combined with appropriate statistical analyses provide a time-efficient and cheap method for carrying out serological surveys to determine the community-burden of RSV.

1.7 Contributed - Data Science: Words

Tuesday 3 September 9am

Feature2Vec: Distributional Semantic Modelling of Human Property Knowledge

Steven Derby[1], Paul Miller[2], Barry Devereux[2]
[1] *Queens University Belfast, ECIT, Department of Data Science and Scalable Computing,* [2] *Queens University Belfast, School of Electronics, Electrical Engineering and Computer Science*

In computational linguistics, the distributional hypothesis of word meaning has allowed us to construct distributional semantic models by mining large corpora of text to extract co-occurrence statistics of words. Neural approaches such as word2vec learn two sets of matrix representations from mined word-context counts, by using dot product between target and context vector with a sigmoid function to measure the probability of positive association. By using negative sampling techniques and gradient descent optimization, we can learn an approximation of word meaning. Our goal is to construct vector representations for a set of human-derived properties by using a neural topology, similar to the skip-gram word2vec, which uses the target word to predict the surrounding windowed context from mined statistics. Surveying a number of concepts for human interpretable features is costly and time-consuming, but unsupervised learning of vector space models from text data is cheap and accessible. We learn feature meaning by sampling a tiny subset of a pretrained set of word embeddings for which we know the properties. Negative sampling along with gradient descent applied only to the matrix of representations allows us to learn feature meaning in relation to the pretrained word vectors. In this case, a word and a feature have meaningful association if their vectors are close together, which we can measure using cosine similarity. Ranking these features for a given concept, we can extract salient features for the word. Furthermore, since these come from a wider vector space model, we can sample unseen words for features. The process allows us to extract possible feature of words, which could make further surveying the concepts for properties much faster. Active learning would then allow us to repeat this process with a larger lexicon which could be then surveyed again, this time with a higher probability of correctly sampling features.

1.7 Contributed - Data Science: Words

Tuesday 3 September 9am

Clustering the citation network of a computer science conference

Clement Lee
Lancaster University

Social network analysis can be applied to a wide range of topics and data. One of its applications is in bibliometrics, where networks concerning academic publications are being constructed and then analysed in a quantitative way. While different networks could be constructed from the same data, it is networks with publications as the nodes that received less attention. In our work, we analysed these citation networks to gain insights into how publications, rather than authors, are grouped .The data concerned are the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. As the premier conference in Human-Computer Interaction (HCI), it has amassed over 6000 publications from 1981 to 2018. These publications are not only electronically available but also properly indexed in the ACM Digital Library (dl.acm.org). The detailed information on the references of these publications allows identification of citations between these ~6000 publications, and therefore the construction of the full citation network with ease.After collecting and cleaning the data, a stochastic block model is applied, which is a popular model in social network analysis for clustering nodes. Bayesian inference, specifically Markov chain Monte Carlo (MCMC), is carried out to estimate the group memberships of the publications, which are treated as latent variables in the model. After modelling and inference, the texts of the publications in each group are processed, to find out the over-represented words, which can be seen as the topic words of the group. It is revealed that these topics found correspond very well with the main themes in the field of HCI. This suggests that citation networks are a good data source, alternative to collaboration networks, that can be explored to understand the landscape of a field in the academic literature.

1.8 Contributed - Medical: Changing practices

Tuesday 3 September 9am

A fresh look at the James-Stein estimator shows that 'dynamic borrowing' of historical data is an illusion

Nicholas Galwey
GlaxoSmithKline

There are strong reasons for seeking to use historical control data to supplement the results from a clinical trial or other experiment, and it has been suggested that this approach is of value even when the only information concerning the bias of the historical data comes from the observed offset between historical and concurrent controls – an approach known as 'dynamic borrowing' (Viele et al., 2014). However, the properties of the James-Stein estimator indicate that this may not be the case. This estimator of the mean of each of p groups of observations can be obtained by shrinking the usual least-squares estimates towards the grand mean, and always gives an increase in the average accuracy of the estimates (i.e. a reduction in mean square error (MSE) = bias + variance), provided that p >= 3. It is intuitively plausible that no such gain in accuracy is possible when p = 2, but the papers that introduced the James-Stein estimator (Stein, 1956; James and Stein, 1961) are heavily mathematical, and for many readers will not provide an easily-understood, rigorous basis for this limitation. This presentation will demonstrate the limitation in terms familiar to statisticians, relating it to the balance between two sources of information on the among-group variance: (mean(group of interest) – mean(other groups)) and var(means of other groups). When p = 2 the second source is absent and the information on bias and variance for the group of interest is confounded, confirming that nothing can be gained from the use of historical control data in this way. A straightforward, transparent way to overcome this limitation is explicitly to place an informative prior distribution on the bias.

References: Viele, K., et al. (2014) Pharmaceutical Statistics 13:41-54.Stein, C. (1956) Proc. Third Berkeley Symp. Math. Statist. Prob. 1:197–206.James, W. and Stein, C. (1961) Proc. Fourth Berkeley Symp. Math. Statist. Prob. 1:361-379.

1.8 Contributed - Medical: Changing practices

Tuesday 3 September 9am

Improving transparency about "power" and trade-offs in subgroup selection: assessing criteria and statistical models for subgroup selection

Claudia Geue[1], Neil Hawkins[1], Jeremy Oakley[2], Richard Grieve[3], Monica Hernandez[2], James Carpenter[3]
[1] *University of Glasgow,* [2] *University of Sheffield,* [3] *London School of Hygiene & Tropical Medicine*

Introduction: Population health can be improved by identifying treatments with favourable risk-benefit or cost-effectiveness ratios for identifiable sets of patients. This is described as precision medicine and underpinned by subgroup analysis. Subgroup analysis requires statistical models that address multiplicity to avoid inflation of type-1 error, which would lead to inappropriate identification of differential treatment effects. However, this should be balanced against the risk of missing true subgroups. We suggest metrics to assess the performance of a given approach for subgroup identification in hypothetical scenarios where true subgroup effects are known. These criteria could be used to describe the "power" of a proposed subgroup analysis or in simulation studies that compare alternative statistical models.

Methods: We reviewed performance metrics in published simulation studies of subgroup analysis. Based on this we suggest a set of metrics that consider estimation and inference and reflect trade-offs between avoiding false positives and false negatives in subgroup analysis.

Results: Subgroup identification methods ranged from t-tests or likelihood-ratio tests including and excluding the predictive term for individual models, to methods using step functions, splines or simultaneous threshold interaction modelling. Performance metrics included the rate of identifying correct treatment-covariate interactions, true-positive rate, true-negative rate, the probability of detecting at least one false-negative subgroup and the probability of detecting at least one false-positive subgroup. Most studies did not mention assessing simulations in terms of power and associated trade-offs. One study assessed model performance under different scenarios for sample size and power.

Conclusion: Misinterpreting results from subgroup analyses could potentially harm patients (missing true subgroups/accepting false subgroups) or lead to futile and costly trials (falsely identified subgroups). Decision-makers must balance the risk of accepting 'spurious' subgroup effects against the risk of rejecting true subgroup effects. Transparency is required in particular about trade-offs associated with avoiding false negatives and false positives.

1.8 Contributed - Medical: Changing practices

Tuesday 3 September 9am

Epidemiological characterisation and classification of disease

Anthony Webster, Benjamin Cairns, Robert Clarke
NDPH, University of Oxford

Modern healthcare classifies diseases by anatomical systems. We used UK Biobank data and their linked Hospital Episodes Statistics to explore whether a combination of Big Data and machine learning methodologies can provide new insights into classification systems for disease. We considered whether supervised learning methods can generate conventional groupings, such as the International Classification of Diseases Version 10 (ICD-10), and explored the potential for novel data-driven classification of diseases. Our aims were to: (i) to provide new insights into the causes and consequences of disease, (ii) to develop new methods to characterise the links between diseases and their underlying risk factors, and (iii) to understand and improve systems for disease classification. By understanding the links between multiple diseases and their underlying risk factors, we hope to transform the way that we study and manage individual diseases and associated multi-morbidity patterns.

2.1 Medical: Investigation of and overcoming practical dilemmas within medical research

Tuesday 3 September 10.10am

Under-representation in Clinical Trials: Participants with rare diseases, reporting and awareness.

Helen McAneney
Queen's University Belfast

Introduction: Trial populations need to reflect those in the community who may benefit from the treatment being tested if the results of the trial are to maximise patient health gains. The European Union defines a rare disease as one affecting fewer than 5 in 10,000 of the general population. With over 8,000 known rare diseases, almost 6% of the population will be affected by a rare disease. This equates to approximately 3.5 million people in the UK and 30 million people across Europe. Consequently one would expect 6% of those recruited to clinical trials to have a rare disease; however this is at best unknown and likely to not be the case. This study will investigate the representation of rare disease participants within clinical trials and possible reasons for under-representation of these participants, including (i) lack of reporting that a participant has a rare disease, (ii) potential exclusion due to the recruitment criteria to the clinical trial and/or (iii) lack of awareness of rare diseases.

Methods: A sample of clinical trials for anti-hypertension medications, listed on ClinicalTrials.gov, will be investigated. Representation of participants will be considered by analysing the reporting of participants' rare disease and the inclusion/exclusion criteria from protocols, reports and publications of the clinical trials. Awareness of rare diseases will be explored through a network analysis of the citation of rare disease literature using CitNetExplorer, specifically citation of the European Union definition of a rare disease.

Potential relevance and impact: The aim of the UK Strategy for Rare Disease is to 'ensure no one gets left behind just because they have a rare disease'. Clinical trials are pivotal to the improvement of patient health, and require representation of all patients, including and inclusion of rare disease patients in clinical trials.

2.2 Official Statistics & Public Policy: International comparisons of health outcomes - opportunities and challenges of using routinely collected administrative databases

Tuesday 3 September 10.10am

Comparing maternal and child outcomes across countries using prospectively planned, pooled analyses of administrative data: the devil is in the detail

Katie Harron
UCL

Careful review and comparison of different systems can help provide insight into the types of system that deliver the best health outcomes for mothers and their children. In particular, international comparisons provide an important opportunity to help us focus on the specifics of care that are likely to lead to improved outcomes. Provided careful consideration is given to the comparability and quality of data collected in different contexts, international comparisons can offer powerful external benchmarks and evidence on where and how improvements to health and services might be made. In this talk I present methods for bringing together data from different countries to facilitate cross country comparisons, based on three international studies relating to infant admissions, outcomes of teenage motherhood, and long term maternal mortality in mothers of infants with neonatal abstinence syndrome. I discuss the strengths of using hospital data captured in different settings and approaches used to address limitations in comparability. Results demonstrate that even when countries appear to have similar healthcare systems, detailed understanding of healthcare seeking behaviours and the policy context is vital in interpreting different patterns of outcomes. Whilst cross country comparisons may not be able to provide definitive answers, they can help us raise the right questions about the best policies to promote maternal and child health.

2.2 Official Statistics & Public Policy: International comparisons of health outcomes - opportunities and challenges of using routinely collected administrative databases

Tuesday 3 September 10.10am

International comparisons of primary care quality using admissions for ambulatory care sensitive conditions: the example of asthma in children

Irina Lut[1], Kate Lewis[1], Linda Wijlaars[1], Tiffany Fitzpatrick[2], Hong Lu[2], Astrid Guttman[2], Sharon Goldfeld[3], Shaoke Lei[3], Geir Gunnlaugsson[4], S Hrafn Jonsson[4], Reli Mechtler[5], Mika Gissler[6], Anders Hjern[7], Pia Hardelid[1]
[1] UCL Great Ormond Street Institute of Child Health, London, UK, [2] ICES & Dalla Lana School of Public Health, Toronto, Canada, [3] Murdoch Children's Research Institute & Department of Paediatrics, University of Melbourne, Australia, [4] School of Social Sciences, University of Iceland, Iceland, [5] Johannes Kepler University, Austria, [6] National Institute for Health and Welfare, Finland, [7] Centre for Health Equality Studies & Department of Clinical Epidemiology, Karolinska Institutet, Sweden

Objectives: International comparisons of health indicators are a powerful tool in health research. Differences in delivery of care mechanisms or how measures are recorded could affect observed differences in health outcomes across countries that have similar health systems and national income levels. We examine challenges in interpreting cross-country comparisons of hospital admission rates for asthma as an indicator of quality of primary care provided to children with long-term conditions.

Methods: Seven jurisdictions provided administrative data on asthma-related hospital admissions and population denominators for children aged 6-15 years inclusive between 2008 and 2015. Five countries provided admissions with asthma as a secondary diagnosis and three included asthma-related attendances to emergency departments. We compared incidence rate ratios (IRRs) between admissions with asthma as the primary diagnosis, admissions with asthma as a secondary diagnosis and asthma-related emergency department attendance.

Results: 74,682 asthma admissions were recorded among 56 million children. Across all jurisdictions, hospital admission rates were higher for boys aged ≤12 and higher for girls 13-15 years. Over our study period, Victoria (Australia) had the highest primary admission rate and Iceland the lowest, with an 8-fold variation between countries. The rate for admissions with asthma recorded as a secondary diagnosis or within emergency departments was over three times higher than for asthma primary diagnosis admissions (IRR 3.5, 95%CI: 3.5-3.6 and IRR 3.0, 95%CI: 3.0-3.1 respectively). England's admission rate with asthma as a primary diagnosis was 7.1 times higher than Sweden (95%CI: 6.7-7.6), but 10.8 times higher when including asthma recorded as a secondary diagnosis (95%CI: 10.4-11.2).

ConclusionsFactors including prevalence and severity of asthma, hospital admission thresholds and recording practices influence hospital admission rates. Clearly defining outcome measures is a critical first step for international comparisons in the quality of care for children with asthma.

2.2 Official Statistics & Public Policy: International comparisons of health outcomes - opportunities and challenges of using routinely collected administrative databases

Tuesday 3 September 10.10am

Using administrative linked datasets to explain differences in child mortality between England and Sweden: opportunities and challenges

Ania Zylbersztejn, Pia Hardelid
UCL Great Ormond Street Institute of Child Health, London, UK

Child mortality is considered an important indicator of health of a nation. England has one of the highest rates of child mortality in Western Europe, nearly twice as high as that of Sweden – a country with comparable levels of economic development and universal healthcare. The differences in child mortality between these two countries have been commonly attributed to differences in quality of healthcare for children and wider socioeconomic inequalities in England. Any differences, however, are likely to be at least partly explained by England's high rates of adverse birth characteristics, such as low birth weight, preterm birth or presence of congenital anomalies. Detailed information about child's health at birth is therefore needed to disentangle the origins of intercountry differences in child mortality.In this presentation, we use national birth cohorts from administrative linked databases in England and Sweden to determine which factors before and after birth contribute most to the gap in mortality between England and Sweden. We used data on almost 4 million singleton live births in England and over 1 million in Sweden in 2003-2012, with detailed information about health of a child at birth (e.g.: length of gestation, birth weight, presence of birth defects), socioeconomic factors and causes and timing of death. We show what steps are needed to assess comparability of national birth cohorts developed from different data sources (hospital admission and mortality records in England, birth, death and hospital admission registers in Sweden), how to compare recording of diagnoses and causes of death across different settings. Our findings and recommendations will be relevant to future comparisons of health outcomes in countries with administrative linked databases.

2.2 Official Statistics & Public Policy: International comparisons of health outcomes - opportunities and challenges of using routinely collected administrative databases

Tuesday 3 September 10.10am

The unique opportunities and critical limitations of using routinely collected national data for international comparisons of maternal and newborn health

Jennifer Zeitlin
Inserm

International comparisons of perinatal health indicators provide performance benchmarks and underpin maternal and child health policies. The Euro-Peristat project began in 2000 with the aim of providing high-quality, comparable health information on maternal and newborn health and healthcare in Europe in order to enable these comparisons. The project compiles data on 30 indicators from participating countries which now include all 28 EU Member States, Iceland, Norway and Switzerland. Participants are clinicians, statisticians and epidemiologists who work with routine data. Indicators are compiled from routine national-level sources, including: vital statistics, birth registers, hospital discharge abstracts and surveys. By conducting research using indicators of perinatal health derived from national routine sources, Euro-Peristat has been able to identify the benefits as well as the dangers of using these data to inform policy and practice. This presentation will illustrate key lessons learned in two main domains (1) defining maternal and infant mortality and morbidity for benchmarking between countries and (2) interpreting correlations between indicators and time periods. These analyses require realistic assessments of data quality limitations and of the appropriate scope of inference to avoid ecological fallacies. These assessments depend on involving representatives from participating countries with knowledge of medical and data collection practices and developing a common vocabulary through collaboration over time. New data on births in 2015 will be used to illustrate potential errors in interpretation as well as the unique insights that can be drawn from these comparisons.

2.3 Environmental / Spatial Statistics: Applications of hidden Markov models in ecology

Tuesday 3 September 10.10am

A continuous-time Arnason-Schwarz model for the annual movement of bottlenose dolphins

Sina Mews[1], Roland Langrock[1], Ruth King[2], Nicola Quick[3]
[1] *Bielefeld University,* [2] *University of Edinburgh,* [3] *Duke University Marine Lab*

Our modelling approach is motivated by individual sighting histories of bottlenose dolphins off the east coast of Scotland. Due to ongoing offshore development, conservation managers seek to better understand the temporal movement patterns of the dolphin population between different sites. The Arnason-Schwarz (AS) model is often used to analyse such transitions between different states, which here correspond to the location (site) of a given individual. Within the AS model, transitions between states are modelled using a discrete-time Markov chain. In our case, however, the capture occasions do not follow a regular sampling protocol, which is why we develop a continuous-time analogue of the AS model. In contrast to the AS model, the new modelling framework does not require capture occasions to be regularly spaced in time.Statistical inference is carried out by regarding the capture-recapture data as realisations from a (continuous-time) hidden Markov model (HMM), where an individual's sighting history corresponds to the observed state-dependent process and the individual’s (true) state corresponds to the unobserved state process. Embedding the capture-recapture setting in the HMM framework allows the associated efficient algorithms to be used in particular for (numerical) maximum likelihood estimation and state decoding. For scenarios with time-varying covariates affecting the state-switching rates, we develop an approximate maximum likelihood approach.In our present analysis, we are particularly interested in seasonal effects on the migration rates of bottlenose dolphins along the Scottish east coast. The results reveal seasonal migration patterns between two main areas, information that can help to inform conservation management. While motivated by a particular data set, our modelling framework is generally applicable to irregularly sampled capture-recapture data where individuals traverse through different states.

2.3 Environmental / Spatial Statistics: Applications of hidden Markov models in ecology

Tuesday 3 September 10.10am

A test for the underlying state-structure of Hidden Markov models: A case study of partially observed capture-recapture data

Rachel McCrea
University of Kent

Hidden Markov models (HMMs) are being widely used in the field of ecological modelling, however determining the number of underlying states in an HMM remains a challenge. Here we focus on a special case partially-observed capture-recapture models, where some animals are observed but it is not possible to ascertain their state, whilst the other animals' states are assigned without error. We propose a mixture test of the underlying state structure generating the partial observations, which assesses whether they are compatible with the set of states directly observed in the capture-recapture experiment. I demonstrate the performance of the test using simulation and through application to a data set of Canada Geese.

2.3 Environmental / Spatial Statistics: Applications of hidden Markov models in ecology

Tuesday 3 September 10.10am

Modelling latent animal movement and behaviour in population abundance surveys using hidden Markov models

Richard Glennie[1], Stephen Buckland[1], Roland Langrock[2], David Borchers[1]
[1] *University of St Andrews,* [2] *Bielefeld University*

Distance sampling and spatial capture-recapture are statistical methods to estimate the number of animals in a wild population based on encounters between these animals and scientific detectors. Both methods estimate the probability an animal is detected during a survey, but do not explicitly model animal movement and behaviour. The primary challenge is that animal movement in these surveys is unobserved; one must average over all possible histories of each individual. In this talk, a general statistical model, with distance sampling and spatial capture-recapture as special cases, is presented that explicitly incorporates animal movement. An efficient algorithm to integrate over all possible movement paths, based on quadrature and hidden Markov modelling, is given to overcome the computational obstacles. For distance sampling, simulation studies and case studies show that incorporating animal movement can reduce the bias in estimated abundance found in conventional models and expand application of distance sampling to surveys that violate the assumption of no animal movement. For spatial capture-recapture, continuous-time encounter records are used to make detailed inference on where animals spend their time during the survey. In surveys conducted in discrete occasions, maximum likelihood models that allow for mobile activity centres are presented to account for transience, dispersal, survival, and heterogeneous space use. These methods provide an alternative when animal movement and behaviour causes bias in standard methods and the opportunity to gain richer inference on how animals move, where they spend their time, and how they interact.

2.3 Environmental / Spatial Statistics: Applications of hidden Markov models in ecology

Tuesday 3 September 10.10am

Modelling population dynamics using hidden Markov models

Takis Besbeas
Athens University Economics Business / University of Kent

Hidden Markov models have gained popularity within the statistical ecology community thanks to their flexibility to accommodate various types of time series data. They have been routinely applied to model animal movement, but their potential has also been demonstrated in capture-recapture and occupancy modelling. In this talk I will discuss recent developments for modelling population dynamics using hidden Markov models. In particular, we show how hidden Markov model methodology provides a flexible and efficient framework for parameter estimation and model selection from time series data of population abundances. We also show how hidden Markov modelling machinery can be used to provide exact inference for Integrated Population Modelling (IPM), where demographic data are also available. In both of these cases, a single hidden Markov model is employed, corresponding to data from a single species. We last show how multiple hidden Markov models can be engaged to provide a Multispecies Biodiversity Indicator when indices of abundance from multiple species are available. We illustrate the methodology using real data in each case.

2.4 Social Statistics: The better understanding of society - methodological innovation on Understanding Society

Tuesday 3 September 10.10am

Understanding the impact of web mode on quantitative analysis of data from Understanding Society

Paul Clarke, Yanchun Bao
University of Essex

Understanding Society has, until recently, mainly used interviewers to administer its questionnaires. However, starting from Wave 8, it is now allowing individuals to participate via the web. Survey mode is known to affect participants' responses so, to assess its impact, a randomized sequential-design experiment was carried out during the first year of fieldwork for Wave 8. This design, akin to encouragement designs in clinical research, permits the identification of mode effects which would otherwise be confounded by non-random selection. We set up a general framework based on extended structural mean models to efficiently estimate the effects of mode not only on the mean but also the variance, covariance and (for categorical variables) the entire distribution of mixed-mode survey variables. We also show how this framework can be used by analysts to test whether the introduction of web mode has made a difference to the results of their particular analysis.

2.4 Social Statistics: The better understanding of society - methodological innovation on Understanding Society

Tuesday 3 September 10.10am

A latent class approach to inequity in health using biomarker data

Apostolos Davillas[1], Vincenzo Carrieri[2], Andrew Jones[3]
[1] *Office for Health Economics and University of Essex,* [2] *University of Catanzaro,* [3] *University of York*

We develop an empirical approach to analyse, measure and decompose Inequality of Opportunity (IOp) in health, based on a latent class model. This addresses the limitations that affect earlier work in this literature concerning the definition of types - such as partial observability, the ad hoc selection of circumstances, the curse of dimensionality and unobserved type-specific heterogeneity - that may lead to either upwardly or downwardly biased estimates of IOp. We apply the latent class approach to quantify IOp in allostatic load, a composite measure of our biomarker data. Using data from Understanding Society, we find that a latent class model with three unobserved types best fits the data and that these types differ in terms of their observed circumstances. Decomposition analysis shows that about two-thirds of the total inequality in allostatic load can be attributed to the direct and indirect contribution of circumstances.

2.4 Social Statistics: The better understanding of society - methodological innovation on Understanding Society

Tuesday 3 September 10.10am

Obtaining consent for the linkage of social media data with large-scale population surveys

Tarek Al Baghal
University of Essex

In the light of issues affecting survey quality, like the increase in unit non-response, it has been argued that social media data from sources like Twitter can be used as a viable alternative. However, there are also a number of shortcomings with Twitter data such as the potential for it being unrepresentative of the wider population, and the inability to validate exactly whose data it is you are collecting. A useful way forward is to combine survey and Twitter data to supplement and improve both, but this requires consent. This study explores these consent decisions. Our findings suggest that consent rates for data linkage are relatively low, and this is in part mediated by survey mode, where face-to-face surveys have higher consent rates than web versions

2.4 Social Statistics: The better understanding of society - methodological innovation on Understanding Society

Tuesday 3 September 10.10am

Do Income Summary Screens Improve Income Data Quality?

Paul Fisher[1], Jonathan Burton[1], Tom Crossley[1], Annette Jackle[1], Alessandra Gaia[2]
[1] *University of Essex,* [2] *University of Milan-Bicocca*

Income data collected as part of household surveys are critical for the study of material living standards. Survey respondents are known to misreport their income, but the types of error are not well-documented. We experiment with the use of an Editable Summary Screen (ESS) during data collection to improve income data quality by a) testing two version of the ESS in a large scale panel survey and b) shed light on the types of reporting errors made by survey participants by classifying the revisions

2.5 Methods and Theory: Modern Fisherian perspectives on inference

Tuesday 3 September 10.10am

Modified maximum likelihood estimation through adjusted scores

Nicola Sartori[1], Michele Lambardi di San Miniato[1], Nancy Reid[2]
[1] *University of Padova,* [2] *University of Toronto*

For regular parametric problems, modified maximum likelihood estimates can be obtained by simple adjustments of the score equation. The choice of the adjustment determines the properties of the resulting estimator. The most notable example is given by Firth's bias reduced estimator. Following the same idea, Kenne Pagui et al. (Biometrika, 2017) proposed a different adjustment that leads to a componentwise median unbiased estimator. Here, we propose a new adjusment that provides an estimator such that each component, or block of components, is asymptotically equivalent to the corresponding estimator obtained from a modified profile likelihood, thus correcting for the presence of the remaining components, treated as nuisance parameters. This estimator partly retains the equivariance property of the maximum likelihood estimator and, as mean and median bias reduced estimators, has the same asymptotic distribution.

2.5 Methods and Theory: Modern Fisherian perspectives on inference

Tuesday 3 September 10.10am

Conditioning and Randomisation in Selective Inference

Alastair Young
Imperial College London

Selective inference is concerned with performing valid statistical inference when the questions being addressed are suggested by examination of data, rather than being specified before data collection. Adjustment for double use of the data through `conditioning on selection' has been proposed as a route to valid inference in a frequentist framework. The Bayesian standpoint on selective inference is less clear, but it may be argued that conditioning on selection is required unless the selection takes place on the parameter space as well as on the sample space. In this talk, we will argue that classical ideas of data randomisation and appropriate conditioning are necessary to yield a theoretically justified and practically appealing Bayesian inference in the selective context.

2.8 Communicating & Teaching Statistics: Recent innovations in school statistics: remediations

Tuesday 3 September 10.10am

Recommendations for assessment of statistics in A level Mathematics

James Nicholson
Durham University

Statistics emerged as a discipline to address pressing practical problems. In the UK, this has not been reflected in school statistics curricula, where students often work with small-scale invented data to develop mastery of statistical technique. Recent curriculum reforms set out to improve this situation; students are expected to work in class with a large authentic data set. However, in the associated high-stakes assessment, there has still been very little emphasis on statistical skills such as interpreting data and drawing conclusions, and a great deal of emphasis on procedural skills, and on factual recall to demonstrate that they have worked with the pre-release data. Apart from questions related to the pre-release large data set for each qualification, there were no questions in any of the specimen papers or the live papers in 2018 and 2019 that use real data.Contexts are often rather banal. One of the reasons for this may be that mathematics does not have any embedded contexts in the way that statistics assessed within Biology, Psychology, Politics, Social Sciences etc. does. Some of the assessment is not fit for purpose and much of it does not allow students to see why statistics is an important and relevant discipline.I will talk briefly about some examples of poor assessment practice and some curriculum issues which may exacerbate the problems with assessment, but the main focus will be the recommendations for assessment of statistics which have been developed by a sub-group convened by the A level Contact Group of the Royal Society's Advisory Committee on Mathematics Education.

2.8 Communicating & Teaching Statistics: Recent innovations in school statistics: remediations

Tuesday 3 September 10.10am

Using real data across the curriculum at secondary level

Darren Macey
Cambridge Maths

Despite the increase in the amount of ‘real data’ which is accessible now through the internet, many teachers find it difficult to locate datasets which are suitable for use with the content of the mathematics curriculum. Although curriculum documents talk aspirationally about the use of technology in statistics, there are few resources provided which stimulate the classroom practice. I am working on developing two types of resources which I hope can help make it easier to use real data in schools, both in mathematics and across other subjects.One is to identify, and make available, real datasets which have some numerical variables and some categorical variables, where the full dataset can be disaggregated into groups by a categorical variable and then summary statistics, or graphical displays can be constructed for the different groups to allow meaningful comparisons. Currently, when summary statistics or graphs of a dataset are constructed, usually nothing more is done – it is an end in itself. These new curated datasets could be used in mathematics to provide the practice required in learning statistical techniques, with an opportunity to develop skills in describing and interpretation that are an increasingly important aspect of the statistical expertise we wish students to develop. They could also be used in other subject areas to which the dataset is relevant.The second is to produce data visualisations which allow students to engage with datasets involving more variables than they usually meet in the mathematics curriculum, but which are commonplace (at least implicitly) in social science and humanities disciplines. I will show an example, based on a visualiser developed by the SMART Centre at Durham, along with curriculum resources which could be used in mathematics classrooms as well as in disciplines where inequalities in educational attainment are explicitly studied.

3.1 Medical: Papers from the Journal of the Royal Statistical Society

Tuesday 3 September 2.10pm

Landmark linear transformation model for dynamic prediction with application to a longitudinal cohort study of chronic disease

Yayuan Zhu[1], Liang Li[2], Xuelin Huang[2]
[1] *University of Western Ontario,* [2] *University of Texas MD Anderson Cancer Center*

Dynamic prediction of the risk of a clinical event by using longitudinally measured biomarkers or other prognostic information is important in clinical practice. We propose a new class of landmark survival models. The model takes the form of a linear transformation model but allows all the model parameters to vary with the landmark time. This model includes many published landmark prediction models as special cases. We propose a unified local linear estimation framework to estimate time varying model parameters. Simulation studies are conducted to evaluate the finite sample performance of the method proposed. We apply the methodology to a data set from the African American Study of Kidney Disease and Hypertension and predict individual patients' risk of an adverse clinical event.

3.1 Medical: Papers from the Journal of the Royal Statistical Society

Tuesday 3 September 2.10pm

Semiparametric Model for Bivariate Survival Data Subject to Biased Sampling

Jin Piao[1], Jing Ning[2], Yu Shen[2]
[1] *University of Southern California,* [2] *The University of Texas MD Anderson Cancer Center, Houston, USA*

To better understand the relationship between patient characteristics and their residual survival after an intermediate event such as the local cancer recurrence, it is of interest to identify patients with the intermediate event and then analyze their residual survival data. One challenge in analyzing such data is that the observed residual survival times tend to be longer than those in the target population, since patients who die before experiencing the intermediate event are excluded from the identified cohort. We propose to jointly model the ordered bivariate survival data using a copula model and appropriately adjusting for the sampling bias. We develop an estimating procedure to simultaneously estimate the parameters for the marginal survival functions and the association parameter in the copula model, and use a two-stage expectation-maximization algorithm. Using empirical process theory, we prove that the estimators have strong consistency and asymptotic normality. We conduct simulations studies to evaluate the finite sample performance of the proposed method. We apply the proposed method to two cohort studies to evaluate the association between patient characteristics and residual survival.

3.1 Medical: Papers from the Journal of the Royal Statistical Society

Tuesday 3 September 2.10pm

Adaptive design in surveys and clinical trials: similarities, differences and opportunities for cross-fertilization

Michael Rosenblum[1], Peter Miller[2], Benjamin Reist[3], Elizabeth Stuart[4], Michael Thieme[2], Thomas Louis[4]

[1] Johns Hopkins University, [2] US Census Bureau, Washington DC, USA, [3] US Department of Agriculture National Agricultural Statistics Service, Washington DC, USA, [4] Johns Hopkins Bloomberg School of Public Health, Baltimore, USA

Adaptive designs involve preplanned rules for modifying an on-going study based onaccruing data. We compare the goals and methods of adaptation for trials and surveys, identifysimilarities and differences, and make recommendations for what types of adaptive approachesfrom one domain have high potential to be useful in the other. For example, clinical trials couldbenefit from recently developed survey methods for monitoring which groups have low responserates and intervening to fix this. Clinical trials may also benefit from more formal identification ofthe target population, and from using paradata (contextual information collected before or duringthe collection of actual outcomes) to predict participant compliance and retention and then tointervene to improve these. Surveys could benefit from stopping rules based on informationmonitoring, applying techniques from sequential multiple-assignment randomized trial designsto improve response rates, prespecifying a formal adaptation protocol and including a datamonitoring committee.We conclude with a discussion of the additional information, infrastructureand statistical analysis methods that are needed when conducting adaptive designs, as well asbenefits and risks of adaptation.

3.3 Environmental / Spatial Statistics: Using electronic health records to model spatial variation in disease risk

Tuesday 3 September 2.10pm

Disaggregation of areal unit count data

Craig Anderson, Kamol Sanittham, Duncan Lee
University of Glasgow

Disease mapping is the field of epidemiology focusing on estimating the spatial pattern of disease risk across a geographical region. Typically the region of interest is subdivided into a set of administrative districts, with the disease data consisting of aggregated disease counts at this district level. This means that inference is also generally restricted to estimating disease risk at the district level.Such inference can, however, be susceptible to the modifiable areal unit problem (MAUP), whereby the estimated risk surface can be affected by the arbitrary choice of district boundaries. In reality, the district-level count is an aggregation of point level disease cases, and it is possible that the results could be completely different if a different spatial partition of the region was selected.In this research, we aim to address this problem by producing “disaggregated” disease risk estimates based on a regular 1km x 1km grid. We consider two main approaches – the first is based on multiple imputation and the second on data augmentation. The method is illustrated using an application to respiratory hospital admissions in Glasgow, Scotland.

3.3 Environmental / Spatial Statistics: Using electronic health records to model spatial variation in disease risk

Tuesday 3 September 2.10pm

Spatio-temporal Modelling of Trends in Benign Prostate Hyperplasia drugs prescriptions in Scotland

Federico Andreis[1], Ashleigh Ward[1], Emanuele Giorgi[2]
[1] *University of Stirling,* [2] *University of Lancaster*

Over the last decade, increased access to official prescription records has provided a wealth of opportunities for the application of advanced statistical methodology in order to better inform public health policies. In this talk, our interest is on modelling the spatial and temporal variation in the prescriptions for the treatment of Benign Prostate Hyperplasia (BPH) symptoms in the Scottish population. We first describe how to extract general practice-level data from the NHS Scotland Open Data portal and then link these to socio-economic indicators (e.g. measures of deprivation and rurality of the patients base). In the second part of the talk, we present a spatio-temporal analysis of the monthly number of daily doses for drugs belonging to two groups (α-1 blockers and 5-α reductase inhibitors) prescribed by GPs in Scotland over a period of 4 years. Our results show an increasing trend in the number of BPH-related prescriptions, which is consistent with both a higher prevalence of prostate-related conditions and an improved detection thereof, likely ascribable, respectively, to the ageing of the population and the increased publicity about prostate cancer screening. While trends are fairly similar across health boards, baseline rates of prescriptions for the two drug groups tend to differ, possibly reflecting local effects due to the influence of the Area Drug and Therapeutics Committees. Lastly, the inspection of the spatial residuals of the model seems to point at the existence of geographic areas (of non-strictly administrative nature) that exhibit similar prescribing behaviours.

3.3 Environmental / Spatial Statistics: Using electronic health records to model spatial variation in disease risk

Tuesday 3 September 2.10pm

A spatially discrete approximation to log-Gaussian Cox processes for modelling spatially aggregated disease counts data

Olatunji Johnson, Peter Diggle, Emanuele Giorgi
Lancaster University

The increased availability of electronic health records (EHRs) opens up new opportunities to study variation in disease risk. However, the use of EHRs for geospatial analysis is hampered by the lack of information on the residence or full postcode of the patients which is due to confidentiality reasons. Instead, the Lower Super Output Area (LSOA) of residence of the patients is usually provided in the UK. This leads to the aggregation of health outcome over such partitions. As variation in disease risk occurs in a spatial continuum irrespective of the format in which the data are available, in this talk, we discuss first that log-Gaussian Cox process (LGCP) can be considered as a natural statistical paradigm for modelling spatially aggregated disease count data, and second discuss the proposed computationally efficient discrete approximation to LGCP model. Furthermore, we discuss an application to COPD emergency admission covering admissions from 2012 to 2018 in South Cumbria and North Lancashire, England, UK. We carry out a spatial prediction of the incidence and unexplained risk of COPD emergency admission. We identify areas with high unexplained risk. And advise that urgent attention should be given to those areas by deploying specialist community COPD nurses as this will help save valuable hospital resources. The proposed methodology is implemented in the open-source R package SDALGCP.

3.5 Methods and Theory: Pseudo likelihood theory and methods

Tuesday 3 September 2.10pm

Robustness of inference for GLMMs under model misspecification

Helen Ogden
University of Southampton

Generalised linear mixed models (GLMMs) are a widely-used class of models, but they make some strong assumptions which are mostly made by convention, and are not easy to verify. It is therefore important to understand the impact on the resulting inference if the true data generating process does not obey these assumptions. I will consider the impact of various types of model misspecification on pseudo-likelihoods and other methods for conducting inference in GLMMs. Often a full joint model for all the data is not needed to specify a pseudo-likelihood, which might rely instead on some low-dimensional marginal distribution or a small number of moments. It may be easier to write down a sensible model for these small components than it is to specify a full joint model, and pseudo-likelihoods are often described as being relatively robust to misspecification of the full joint model, compared with the full likelihood. I will evaluate the extent to which this is true, using GLMMs as an example model class.

3.5 Methods and Theory: Pseudo likelihood theory and methods

Tuesday 3 September 2.10pm

Towards fully-efficient estimation

Ioannis Kosmidis
University of Warwick

The inverse of the Fisher information matrix in a likelihood problem is i) the variance-covariance matrix of the asymptotic distribution of the maximum likelihood (ML) estimator; ii) the dominant term in the expansion of the finite-sample variance of the ML estimator; and iii) the "lowest" achievable variance-covariance that an unbiased estimator can achieve. "Lowest" here is used to indicate that the difference of the inverse Fisher information from the variance of any unbiased estimator is a positive definite matrix. These three characterizations and the asymptotic unbiasedness of the ML estimator are key justifications for the wide-spread use of the latter in statistical practice. For example, standard regression software typically reports the ML estimates alongside with estimated standard errors coming from the inversion of the Fisher information matrix at the estimates. Nevertheless, the use of that pair of estimates and estimated standard errors for inference implicitly assumes, amongst other things, that the information about the parameters in the sample is large enough for the estimator to be almost unbiased and its variance to be well-approximated by the inverse of the Fisher information matrix. In this talk, we present results from work-in-progress on a novel estimation framework that aims to bridge the finite-sample gap between estimates and the estimated variance-covariance matrix. We also show results from inferential settings that are well-used in statistical practice.

3.8 Business, Industry & Finance: Contemporary Challenges in Industrial Statistics

Tuesday 3 September 2.10pm

Next generation methods for industrially-focussed earth-observation imagery

Louise Lloyd
Rezatec

Rezatec combines data science with satellite imagery and geo-spatial data to deliver cloud-based analytics to global customers owning and operating land-based assets.The new and growing Data Science team at Rezatec investigates and applies a large range of models to satellite and geo-spatial data. Current work includes identifying different crops and tree species, predicting water leaks and forecasting global commodity harvests a year in advance.This presentation will focus on the world of forestry, and in particular, the massive forests in the USA and Canada. These are too big to observe regularly from the ground, but we can observe them from space. Using freely available high-resolution satellite imagery and some ground data for training, we can identify the number, species and volume of trees across large areas of forest with an accuracy greater than 80%. Typically, these results are achievable with some straightforward machine learning algorithms, but accuracies above this are hard to achieve and there are some situations where the algorithms fail completely. Our ‘next-generation techniques’ combine these machine learning algorithms with statistical and ecological models. They encompass knowledge about plant communities and tree habitats in the hope of achieving interpretable models and accuracies greater than our current levels.

3.8 Business, Industry & Finance: Contemporary Challenges in Industrial Statistics

Tuesday 3 September 2.10pm

Novel methods for senor-based streaming data

Idris Eckley
Lancaster University

The advent of low-cost sensors in the business and industrial landscape has resulted in easy access to high-quality data streams. These streams, if appropriately harnessed, can provide valuable insights into previously unseen operational aspects, and the opportunity to unlock, automate and transform the management of key processes. In parallel, such developments can also result in the need to develop new statistical methodology, for example in the areas of time series, changepoint analysis and anomaly detection. This talk will provide an introduction to the topic, introducing various case studies that highlight recently developed work in this area.

3.8 Business, Industry & Finance: Contemporary Challenges in Industrial Statistics

Tuesday 3 September 2.10pm

Personalised marketing: challenges and recent advances

Arnoldo Frigessi
University of Oslo

Marketing is personalised when every user/customer is approached in a unique way, in an effort to present products/items/services which are of special interest for exactly her/him. Representing and understanding the personal preferences of each unique user, on the basis of highly incomplete data, is the first important challenge. Inference exploits patterns of similarity between users. The second challenge is to perform prediction of the user's actions, in order to optimise personalised marketing. We will describe two projects, one in recommender systems (where accuracy of prediction and diversity of offers are important) and social-network based marketing (where user preferences spread like infections over the network).

PD2 Professional Development: Data FAIRification using R/Rstudio workflows

Tuesday 3 September 2.10pm

Data FAIRification using R/RStudio workflows

Brendan Palmer, Darren Dahly
HRB Clinical Research Facility Cork

Research data are often lost or discarded at the end of a study, despite their continued value. Funders are cognisant of this loss, so researchers are increasingly being asked to prepare their data according to the FAIR Guiding Principles (https://doi.org/10.1038/sdata.2016.18). For many researchers it is unclear how such initiatives can be incorporated into their current day-to-day workflows. The overriding goal of this workshop is to provide attendees with a reproducible workflow based on R/RStudio that adheres to the core FAIR principles of findability, accessibility, interoperability and reuse.

The workshop will be organised and delivered by Drs Dahly and Palmer of the HRB CRF-C, School of Public Health, University College Cork. Drawing on their experience providing statistical support for clinical investigators, and recognising the increased requirements for FAIR data, they have developed an R/RStudio workflow, that runs from project initiation though to automated report generation. This workflow also includes steps to ensure data are ‘FAIRified’ as the workflow culminates in downstream application of globally resolvable unique, persistent identifiers (e.g. DOI) and resultant long term preservation.All materials will be made freely available online in advance of the session. Participants will be invited to follow the instructors on their own laptops as we take a ‘raw’ sample data set (in .csv format) and subsequently clean, ‘FAIRify’ and publish the data object.

The workshop will be broken into four main sections:Introduction to open science and FAIR data stewardship.Building research projects using a defined folder structure coupled with R-projects.'FAIRification' of sample data as part of the data cleaning processPersistent DOI assignment.

The main learning outcome from this session will be a reproducible workflow that attendees can take back to their current workplace and use as a first step towards data ‘FAIRification’ of their own project outputs.

Rapid Fire Talks 1

Tuesday 3 September 3.40pm

A Bayesian Model Averaging Approach to g-Parameter Priors Elicitation

Saheed Afolabi
Ibadan, Nigeria

A special technique that measures the uncertainties embedded in model selection processes is Bayesian Model Averaging (BMA) which depend on the appropriate choices of model and parameter priors. As important as parameter priors' specification in BMA, 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). Therefore, this research aimed at eliciting a modified g-parameter priors to improve the performance of the PMP and predictive ability of the model. From the functional form of the g-priors used; the tools of BMA like Bayes Theorem, Bayes Factor (BF), Posterior Model Probability (PMP), Prior Inclusion Probability (PIP) and Shrinkage Factor (SF) through the modified g-parameter priors gj = established the superiority of the consistency's conditions and asymptotic properties of the prior(s) using the Fernandez, Ley and Steel (FLS) models (1 & 2); and respectively with as sample sizes. The result from the analysis revealed 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. For the three modified g-parameter priors, the best reliability occurred when n = 100; 000 for Model 1 and Model 2 with (0.0631, 0.0521 and 0.0411) and (0.00, 0.00 and 0.00) respectively; also, the best convergence occurred with (0.9343, 0.9460 and 0.9577) and (1.00, 1.00 and 1.00) for Model 1 and Model 2 respectively when n = 100; 000. 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.

Rapid Fire Talks 1

Tuesday 3 September 3.40pm

How GSK is helping to increase statistical capabilities in sub-Saharan Africa

Lindsay Kendall, Annie Stylianou, Nandita Biswas, Agbor Ako
GSK

The demand for qualified experienced statisticians in sub-Saharan Africa is very high, growing, and outnumbers the supply. In response GlaxoSmithKline (GSK) has committed to help increase statistical capabilities on the continent as part of the Africa Non-Communicable Disease (NCD) Open Lab.The Africa NCD Open Lab was established 5-years ago as part of a series of GSK's strategic investments to provide sustainable support for scientific research in the field of NCDs in sub-Saharan Africa. We are currently working in partnership with funders, researchers and academic groups to share expertise and resources to increase the scientific understanding behind the unique attributes of NCDs in the African population. How are GSK statisticians involved?We currently provide statistical support to 19 projects receiving GSK funding and co-funding with Medical Research Council (MRC) UK and MRC South Africa. Support ranges from consulting with the project teams on design issues, approaches to minimise bias, sample size calculations, protocol development and analysis considerations.We are currently providing both funding and support for 6 PhD students, 5 MSc students (another 5 planned for 2019), 3 internships (another 3 planned for 2019) and 1 fellowship (another 2 planned for 2019 and 2020) in collaboration with the sub-Saharan Africa Consortium for Advanced Biostatistics Training (S2ACABT), the London School of Hygiene and Tropical Medicine (LSHTM) and the African Institute for Mathematical Sciences (AIMS).We hosted a week-long workshop in 2017 for 18 African statisticians at GSK's Research and Development facilities in the UK. We rolled out focused statistical, research and NCD-specific training sessions - and provided an opportunity for networking.Other initiatives are also in development!So, what's next?We plan to continue the current statistical support GSK provides in sub-Saharan Africa. Moving forward GSK welcomes suggestions, partnership and collaborations!

Rapid Fire Talks 1

Tuesday 3 September 3.40pm

Valuation of preference-based measures: could borrowing strength from existing countries' valuations produce better estimates

Samer Kharroubi
American University of Beirut

Background: Valuations of preference-based measure such as EQ-5D and/or SF6D have been conducted in different countries. There is a scope of borrowing strength from existing countries' valuations to generate better representative utility estimates. This is explored using two case studies modelling UK data alongside Japan samples to generate Japan estimates.

Methods: Data from two SF-6D valuation studies were analyzed where, using similar standard gamble protocols, values for 241 and 249 states were devised from representative samples of Japan and UK general adult populations, respectively. Two nonparametric Bayesian models were applied to estimate a Japan value set, where the UK results were used as informative priors in the first model and subsets of the Japan dataset for 25 and 50 health states were modelled alongside the full UK dataset in the second. Generated estimates were compared to a Japan value set estimated using Japan values alone using different prediction criterion.

Results: The results allowed the UK data to provide significant prior information to the Japan analysis by generating better estimates than using Japan data alone. Also, using Japan data elicited for 50 health states alongside the existing UK data produces roughly similar predicted valuations as the Japan data by itself.

Conclusion: The promising results suggest that the existing preference data could be combined with data from a valuation study in a new country to generate preference weights, thus making own country value sets more achievable for low-middle income countries. Further research and application to other countries and preference-based measures are encouraged.

Rapid Fire Talks 1

Tuesday 3 September 3.40pm

Statistical reproducibility for (multiple) pairwise tests in pharmaceutical product development

Andrea Simkus
Durham University / AstraZeneca

In pharmaceutical discovery and development, current decisions of statistical tests play an important role. Statistical reproducibility is another property, like power, that provides valuable information and extends our decision-making capacity. We concentrate on a real world test scenario, which employs two pairwise tests: the t-test and its nonparametric counterpart, the Wilcoxon Mann-Whitney test. The scenario involves 6 test groups, which are given an increasing dosage of a drug and a chosen observation is subsequently recorded. The aim of the test is to decide what dosage of the drug is the most effective one. The main motivation is the study of this scenario and its reproducibility. Our aim is to answer the question: Would a repeat of the experiment lead to the same test decision? We employ nonparametric predictive inference (NPI) for reproducibility for the two pairwise tests. The study of reproducibility of the t-test is a novelty. We also consider reproducibility for the overall decision. We adopt NPI bootstrap for calculating reproducibility of both tests as opposed to calculating exact lower and upper probabilities. NPI bootstrap is an alternative to standard bootstrap and it is based on a repeated application of Hill's assumption A(n). We present two approaches. Firstly, we look at separate pairwise comparisons and their reproducibility by creating NPI bootstraps for both test groups, applying the pairwise comparison on those and counting how many times we get the same test decision. Secondly, we study the reproducibility of the decision for the actual test scenario when we carry out multiple comparisons, again by using NPI bootstrap. Among others, we conclude that in this scenario both pairwise tests lead to similar reproducibility. When we move from doing separate to multiple pairwise comparisons, reproducibility probability goes significantly down. We also examine different ways of presenting statistical reproducibility and we introduce them.

Rapid Fire Talks 1

Tuesday 3 September 3.40pm

Assessing local chlamydia screening performance by combining survey and administrative data to account for differences in local population characteristics

Nathan Green[1], Ellie Sherrard-Smith[1], Clare Tanton[2], Pam Sonnenberg[3], Catherine Mercer[3], Peter White[1]

[1] Imperial College London, [2] LSHTM, [3] University College London, Imperial College London

Reducing health inequalities requires improved understanding of the causes of variation. Local-level variation reflects differences in local population characteristics and health system performance. Identifying low- and high-performing localities allows investigation into these differences. We used Multilevel Regression with Post-stratification (MRP) to synthesise data from multiple sources, using chlamydia testing as our example. We used national probability survey data to identify individual-level characteristics associated with chlamydia testing and combined this with local-level census data to calculate expected levels of testing in each local authority (LA) in England, allowing us to identify LAs where observed chlamydia testing rates were lower or higher than expected, given population characteristics. Taking account of multiple covariates, including age, sex, ethnicity, student and cohabiting status, 5.4% and 3.5% of LAs had testing rates higher than expected for 95% and 99% posterior credible intervals, respectively; 60.9% and 50.8% had rates lower than expected. Residual differences between observed and MRP expected values were smallest for LAs with large proportions of non-white ethnic populations. London boroughs that were markedly different from expected MRP values (90% posterior exceedance probability) had actively targeted risk groups. This type of synthesis allows more refined inferences to be made at small-area levels than previously feasible.

Rapid Fire Talks 1

Tuesday 3 September 3.40pm

Counterfactual Analysis Using Censored Duration Data

Andres Garcia-Suaza[1], Miguel Delgado[2]
[1] *Universidad EIA,* [2] *Universidad Carlos III*

We propose standardization techniques for the duration distribution in a population with respect to another taken as standard using right censored data, which forms a basis for counterfactual comparisons between distributional features of interest. Alternative standardizations are based on either a proportional hazard semiparametric specification or a nonparametric specification of the underlying conditional distribution. Applications to the restricted mean survival time and the hazard rate are discussed in detail. The proposal is applied to the counterfactual analysis of spells of unemployment duration gender gaps in Spain between 2004-2007. The behavior in small samples is investigated using Monte Carlo experiments.

Rapid Fire Talks 1

Tuesday 3 September 3.40pm

Drug utilisation reporting using administrative claims data in Ireland

Lea Trela-Larsen[1], Cara Usher[2], Laura McCullagh[2], Michael Barry[2], Cathal Walsh[1]
[1] *University of Limerick,* [2] *National Centre for Pharmacoeconomics, Ireland*

Introduction: In Ireland the National Centre for Pharmacoeconomics (NCPE) assess evidence about the health-benefit, cost-effectiveness (value for money) and budget impact of new drugs compared to currently available treatments. These assessments inform funding decisions for new medicines. Currently there is a lack of assurance post-reimbursement that the predictions made in these assessments were realistic. Analysis of national administrative pharmacy claims data could provide an important source of evidence on how these drugs are used in practice. This evidence would be useful to inform NCPE assessments and the review of funding decisions for these drugs. We aimed to engage with stakeholders and elicit from them the information they deem necessary to fulfil their roles in the drug reimbursement process. This engagement informed the development of drug utilisation reports.

Methods: Stakeholders completed an initial questionnaire to establish their role and information requirements. Their responses helped us to develop our analyses and reports. Draft drug utilisation reports, based on the analysis of national pharmacy claims data, were developed using R. Stakeholders were then asked to provide feedback on these draft reports.

Results: Stakeholders rated information on expenditure (9.2), health outcomes (9) and patient numbers (8) as the highest for importance on a scale from 0 ‘not at all important’ to 10 ‘very important’. Draft reports included information on patient prevalence and incidence over time, patient demographics, mean dose and treatment duration. Stakeholders found the initial drafts useful and clear as high-level reports, but asked for more detail in certain areas.

Conclusions: Stakeholders saw drug utilisation evidence as an important additional source of data to inform decision making. Pharmacy claims data can provide information on expenditure and patient numbers, but not health outcomes. Data linkage to patient registries could provide information on health outcomes. We will seek further feedback from stakeholders to assess how they use our reports.

Rapid Fire Talks 2

Tuesday 3 September 3.40pm

Calculation of Relative Threshold Levels for the Capacity of Benefits from the Arthroplasty Surgery using the Quantile Polynomial Regressions

Sujin Kang[1], Jonathan Cook[2], Andrew Price[3]
[1] *Imperial College London,* [2] *University of Oxford,* [3] *Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, , , , , , ,*

Objective: To calculate threshold levels of the pre-operative patient-reported scores that patients could have benefit by estimating centiles of the study population.

Method: The polynomial based quantile regression were used to estimate an individual's pre-operative patient-reported score based on their probability of gaining a meaningful improvement after hip or knee replacement. The fractional polynomial logistic regression were fitted additionally to examine the benefit of the baseline covariates. Participants: 209,761 patients who underwent hip replacement, and 222,933 patients who underwent knee replacement. Primary outcome measures were the change (post – pre-operative) of the Oxford Knee Score (OKS) and the Oxford Hip Score (OHS), respectively.

Outcomes: The 3rd degree polynomial based quantile regressions to the probability of achieving a meaningful improvement were estimated and validated. The upper thresholds were estimated as being a pre-operative score of 40 for hip and 41 for knee (48 is the maximum score). The predicted (i.e., relative) threshold levels from the 20th quantile regressions (i.e., 80% population coverage) were 25 for the OKS and 32 for the OHS; they showed high sensitivity and poor specificity compared to the threshold levels from the 10th, 30th and 50th quantile regressions. The predicted threshold levels from the 50th quantile regressions (i.e., 50% population coverage) were 36 for the OKS and 38 for the OHS. As the threshold is reduced, the probability of meaningful improvement increased towards maximum probability values.

Conclusion: The polynomial based advanced modeling approaches have been developed using the large routine patient-reported data; it allowed the pre-operative Oxford scores to be used as a guide for referral. The models showed nicely differing proportions of individuals and their expected outcomes with good performance, but also implied the uncertainty about where an individual patient will end up.

Rapid Fire Talks 2

Tuesday 3 September 3.40pm

Mapping the uncertain future of longevity: an ensemble approach for forecasting mortality

Mark Hancock, Guangquan Li, Pete Philipson
Northumbria University

Life expectancy has been on the rise in most countries due to the continuous development in healthcare over the past century. This is positive. However, with the rising average age of human, current plans for pensions and healthcare may need to be revised to remain affordable for a country. To inform appropriate changes to these plans, mortality forecasts need to be reliable with various sources of uncertainty incorporated. The goal of this project is to develop a model ensemble approach to forecast human mortality. Based on Bayesian model averaging (Hoeting et al. 1999), our approach forecasts mortality by probabilistically combining the forecasts from a suite of models so that model uncertainty is accounted for. In a conventional single-model forecasting approach, this source of uncertainty is ignored. Here, we propose a mixture modelling approach to estimate the posterior model probabilities, which are then used as the model weights. Each forecast model contributes towards the final forecasts whilst the estimated model weights quantify the relative contributions of these model.A feature of the proposed mixture approach is that the fitting of the individual models and the estimation of their model weights are carried out within a single fitting. This contrasts the two-stage approach proposed by Kontis et al. (2017) where the model weights are estimated through a cross validation setting. In a situation where there are only a limited number of time points with data available, the single fitting feature from the mixture approach becomes beneficial.To compare the forecast performance of the mixture and the two-stage approaches, a simulation study is performed, assessing the performance of both point and interval forecasts. Our findings show promising results for the mixture modelling approach, whilst both BMA techniques are shown to perform more reliably than using a single-model approach.

Rapid Fire Talks 2

Tuesday 3 September 3.40pm

Multivariate Correction for Attenuation of a Congeneric Measurement Model without Correlated Errors: A Test of the Bock and Petersen Approach

Scott Colwell
University of Guelph

Researchers often employ multi-item measures to test complex theoretical relationships among constructs of interest. These items are often estimated as a congeneric measurement model where items load on to their respective latent variables with differing degrees of precision and error. However, when testing the relationships among the constructs, latent variables are often replaced with composite variables to ease computation. Psychometric research shows that measurement error—the amount of observed variance that cannot be statistically attributed to the construct or known source of measurement error—causes correlations to be attenuated.While attenuation in correlations due to measurement error might be less problematic if all correlations in a network were attenuated to the same degree, observed correlations are usually based on different underlying relationships with different levels of measurement error variance. The result is that correlations are usually attenuated to varying degrees, thereby distorting the observed pattern of relationships relative to the underlying pattern of relationships. This becomes a formidable problem when drawing conclusions about patterns of relationships between constructs that are not influenced by distorting effects of psychometric artifacts such as measurement error variance. Spearman (1904) established a bivariate correction for attenuation between two variables with uncorrelated errors. However, Bock and Petersen (1975) proposed that this approach is limited to only the bivariate condition in that in the multivariate context, it may produce an indefinite correlation matrix that is not admissible for modeling the between construct relationships. While Bock and Petersen (1975) derived a multivariate correction for attenuation, it has yet to be tested in any extensive ways. As a result, this research employs a Monte Carlo simulation to test the Bock and Petersen approach with a congeneric measurement model without correlated errors. Results show the Bock and Petersen approach limits the amount of relative bias in the correlation estimates and furthermore ensures that the resulting correlation matrix is at least semi-positive definite.

Rapid Fire Talks 2

Tuesday 3 September 3.40pm

On Parameter Estimation of the Hidden Gaussian Process in perturbed SDE.

Li Zhou[1], Yury Kutoyants[2]
[1] *Shandong University, Weihai,* [2] *Le Mans University, Le Mans, France*

We present results on parameter estimation and non-parameter estimation of the linear partially observed Gaussian system of stochastic differential equations. We propose new one-step estimators which have the same asymptotic properties as the MLE, but much more simple to calculate, the estimators are so-called "estimator-processes". The construction of the estimators is based on the equations of Kalman-Bucy filtration and the asymptotic corresponds to the small noises in the observations and state (hidden process) equations. We propose conditions which provide the consistency and asymptotic normality and asymptotic efficiency of the estimators.

Rapid Fire Talks 2

Tuesday 3 September 3.40pm

Model Averaging in a Multiplicative Heteroscedastic Model

Alan Wan
City University of Hong Kong

This paper develops a model averaging method for combining maximum likelihood estimators of unknown parameters in a multiplicative heteroscedastic model. A weight choice criterion based on minimising a plug-in estimator of the squared prediction risk of the model average estimator is developed, and the asymptotic optimality of the resultant estimator is established. Our estimator takes explicit account of the model of heteroscedasticity for the error process as well as the uncertainty about the regressor choice in both the mean and variance functions of the model, whereas all existing studies of model averaging are based on estimators that either ignore heteroscedasticity, or assume there is no uncertainty governing the regressor choice in the variance function. Results of our simulation and real data analysis show that more efficient estimates than those produced by heteroscedasticity-robust FMA estimators can be obtained using this new estimator, which also performs better than the traditional AIC and BIC model selection estimators.In recent years, statisticians have gravitated to the use of heteroscedasticity-robust standard errors for the OLS estimator as an outgrowth of the popularity of this technqiue. However, as emphasised by Leamer (2010), this does not necessarily imply that one should no longer model heteroscedasticity because by ignoring heteroscedasticity one can potentially forego substantial gains in estimator's efficiency. As discussed in the Introduction, possible efficiency loss if heteroscedasticity is modeled incorrectly is one major reason for statisticians to shy away from modeling heteroscedasticity, and instead favour the simpler solution offered by the robust standard errors approach.Model averaging offers a way around the problem of model mis-specification by hedging estimates against the very bad models. For this reason, model averaging habitually delivers more accurate estimates of the unknown parameters in finite samples and should be viewed as an alternative to the present status quo of using the less efficient OLS estimator combined with heteroscedasticity-robust standard errors .

Rapid Fire Talks 2

Tuesday 3 September 3.40pm

Robustness of Space-Filling Orthogonal Array Based Composite Design to Missing Observation

Abimibola Oladugba, Ezechukwu Okeke
Department of Statistics, University of Nigeria, Nsukka

Robustness of Space-Filling Orthogonal Array Based Composite Design to Missing ObservationAbimibola Victoria Oladugba1 & Ezechukwu Fidelis Okeke21,2 Department of Statistics, University of Nigeria, NsukkaMissing observation occurs whenever a valid observation is not available for anyone of the experimental unit. In this work, a new space-filling orthogonal-array based composite minimax loss designs which are robust to one missing observation and also enable parameters to be estimated without loss of efficiency were proposed. The designs were constructed using a two-level factorial design and a three-level orthogonal array with space-filling properties such as uniformity and maximin distance. The proposed designs and other composite designs such as space-filling orthogonal-array based composite designs, orthogonal array composite designs, small composite designs and central composite designs were compared based on D-efficiency and precision of regression coefficient estimates by calculating the generalized scaled standard deviations for full model, linear, quadratic, and bilinear terms respectively. It was observed that space-filling orthogonal-array based composite minimax loss designs performed better than other composite designs for all cases.

Rapid Fire Talks 2

Tuesday 3 September 3.40pm

Easy and intuitive multivariate normal quadrant probabilities

Julia Crook
Mayo Clinic

There are well-established formulae for the calculation of multivariate normal quadrant probabilities in the cases of 2 and 3 dimensions. We re-arrange these formulae to be in a simpler and more intuitive form that is easier to remember and implement. We do this using a transformed version of the correlation: $\rho^* = \rho^*(\rho) = \arcsin(\rho) /(\pi/2)$. This transformation is monotonic and such that ρ^* takes slightly attenuated but similar values to ρ: values of ρ^* , like ρ, are in the range -1 to 1 and are equal to ρ when it is 0, 1 or -1. Two other special cases that are easy to remember are: $\rho^* (1/2) = 1/3$ and $\rho^* (-1/2) = -1/3$. The simplified formulae are, with usual notation: $P(Z1>0, Z2>0)= 1/4 (1 + \rho12^*)$, and $P(Z1>0, Z2>0, Z3>0) = 1/8 (1+ \rho12^*+ \rho13^*+ \rho23^*)$. For 4 dimensions, we can extend this to create an exact formula that applies when one of the pairwise correlations is zero. For higher dimensions we have similar exact formula that also require some of the correlations to be zero. These formulae are intuitive and appealing, and subsequently easy to remember and apply.In addition to the situations where quadrant probabilities can be obtained exactly, they can also be useful for rule-of-thumb or approximate calculations. For example, from the above it is easy to see that if Z1 and Z2 are bivariate normal with mean zero and correlation 0.5 then $P(Z1>0, Z2>0) = 1/3$ and hence that $P(Z1>0, Z2<0) = 1/6$. Similarly for a trivariate normal where all three pairwise correlations are 1/3, we have $P(Z1>0, Z 2>0, Z3>0) = 1/4$ along with $P(Z1>0, Z2>0, Z3<0) = 1/12$. The simple formulae, as well as aspects of their derivation, may be particularly useful in a teaching setting.

Rapid Fire Talks 3

Tuesday 3 September 3.40pm

Program evaluation and causal inference for distributional and functional data: estimation of the effects of retirement on health outcomes

Andrej Srakar
Institute for Economic Research (IER), Ljubljana and Faculty of Economics, University of Ljubljana

Statistical analysis of complex, i.e. non-standard data is gaining ground. Analysis of compositions, intervals, histograms, distributions and functions has become more and more common in contemporary statistics and econometrics. Despite several types of regressions existing for symbolic data (e.g. Billard and Diday, 2002; 2006; Dias and Brito, 2015; Irpino and Verde, 2015), causal inference has not been studied so far adequately here. Furthemore, only slowly is it gaining ground using functional data (Zhao et al., 2018). We develop a statistical theory for using instrumental variables with symbolic distributional data, related to the prevailing usage of regressions in such situations, the so-called "two components" Irpino and Verde model (Irpino and Verde, 2015). We apply the findings to a pressing problem in the analysis of the aging process: the effects of retirement on health outcomes. Charles (2004), Neuman (2008), Latif (2013), Insler (2014) and Eibich (2015) conclude that retirement may lead to significant health improvements, but other studies find negative retirement effects (e.g. Dave et al., 2008; Behncke, 2012). We use a panel dataset of Survey of Health, Ageing and Retirement in Europe (SHARE) in Waves 1-6. To address reverse causality in the relationship of retirement and health behaviours we use as an instrument changes in retirement age (Komp, 2017). A novelty in the approach is that we treat countries as units and the variables are aggregated over countries. In this manner, we estimate the effect of the distribution of retirement across countries on distribution of health outcomes over countries (the instrumental variable is distributional as well). We extend the analysis for functional data and aggregate the variables to nonparametric functions, later used in functional linear models. As program evaluation and causal inference has so far not been studied with distributional data (and very seldom with functional data) the article is a significant step ahead in the analysis of complex data.

Rapid Fire Talks 3

Tuesday 3 September 3.40pm

Estimating the correlation between bivariate survival endpoints with semi-competing risks

Yinghui Wei, Lexy Sorrell, Małgorzata Wojtyś
University of Plymouth

Our objectives are to estimate the correlation between bivariate semi-competing risk time-to-event endpoints. Traditional methods to calculate the correlation cannot be used due to the common censoring of time-to-event endpoints. We develop a copula-based approach to estimate the association between semi-competing risk endpoints, where a terminal endpoint can censor the non-terminal endpoint but not vice-versa. Copula based likelihood functions are derived to describe the survival outcomes and the association between them. We estimate the association parameter of a copula and the hazard functions of the marginal distributions, subsequently transform the association parameter to a correlation coefficient. We use several copulas to estimate the association parameter between endpoints, and use information criteria to choose copula models. The proposed methods are applied to a clinical data set and evaluated by an extensive simulation study.

Rapid Fire Talks 3

Tuesday 3 September 3.40pm

Joint modelling of multiple primary outcomes in clinical trials with missing data.

Victoria Vickerstaff, Gareth Ambler, Rumana Z. Omar
University College London

Background: In clinical trials, sometimes multiple primary outcomes are specified. The outcomes maybe of different types, say a mix of survival and continuous outcomes. For example, in a trial investigating the effect of a health intervention on cannabis users, the primary outcomes may be the time to psychiatric relapse and the level of cannabis in the urine. These outcomes are typically associated. Joint models can be used to link survival-type outcomes with continuous outcomes and could provide better insights into the intervention effect. The survival and continuous outcomes may be analysed using a survival submodel and a longitudinal submodel, respectively. These submodels can be linked. One approach to link the submodels is to share parameters between them. Another approach is to use joint random effects.

Objectives: This study evaluates the performance of joint models in terms of bias and efficiency of the estimated treatment effects. The results are compared to the estimates obtained when analysing the outcomes separately.

Methods: Several simulation scenarios were investigated by varying the strength of the association between the outcomes and the level of missing data. Joint models which share parameters or have joint random effects were implemented using the R packages: JoineR, jointModel and FrailtyPack.

Results: The results show that when the outcomes are analysed separately, parameter estimation for the survival outcome is typically biased. The bias is reduced when using the joint models. The joint models had increased standard errors for the estimated treatment effect on the survival outcome compared to analysing the outcomes separately. When there is strong association between the outcomes, the joint random effects models and the model that utilised shared parameters between the longitudinal submodel and the survival submodel performed best in terms of the mean square error of the estimated intervention effects on the survival outcome.

Rapid Fire Talks 3

Tuesday 3 September 3.40pm

Comparing methods of defining hyposmia in a large incident cohort of patients with Parkinson's disease

Sofia Kanavou[1], Michael Lawton[1], Vanessa Pitz[2], Yoav Ben-Shlomo[1], Donald Grosset[2]
[1] *University of Bristol,* [2] *University of Glasgow*

OBJECTIVE: Impaired olfaction (hyposmia) is a common feature in patients with Parkinson's disease (PD), yet there is no standard method to define it. A comparison of four published methods of defining hyposmia using the Tracking Parkinson's cohort, could explain why this is difficult to be established.

METHODS: The 40-item University of Pennsylvania Smell Identification Test (UPSIT) or Sniffin' Sticks (Sniffin) were used to measure olfaction but harmonised, as cut-points for Sniffin were converted from UPSIT values using the Item Response Theory method. The proportion of hyposmic patients was calculated using 4 previously reported definitions:1) Scores ≤ 23 for those aged ≥60 years and ≤ 28 for those <60; 2) Severity grading absolute values, using gender-corrected cut-points; 3) Scores ≤15th centile for normals, age and gender-corrected; and 4) scores ≤15th centile for normals, uncorrected. Control group with age and gender matched to our PD cohort was simulated from published normative data. Between method agreement was measured by Cohen's kappa and Gwet's AC1.

RESULTS: The proportion classified as hyposmic varied between 75.3-98.3% for 1,674 recently diagnosed PD patients and 14.5-52.8% in the simulated healthy group, depending on method. The level of agreement varied significantly: Cohen's kappa ranged from 0.10 to 0.65. Gwet's AC1 ranged from 0.7 to 0.85 and was often substantially higher than the kappa: 1) 0.74>0.11 (Method 1 vs Method 2); 2) 0.85>0.19 (Method 2 vs Method 4). Observed concordance matches better the agreement levels of Gwet's AC1. It estimates agreement probability more accurately, whereas kappa is affected by the prevalence of hyposmia and produces more biased results. The method of defining hyposmia should be carefully chosen when setting research questions and comparing results across different studies of PD.

Rapid Fire Talks 3

Tuesday 3 September 3.40pm

Assessment of alternate data-sources to meet the challenges of timelier and accurate registration of fact-of-death, as needed by record-linkage studies.

Paula Curnow[1], Charlotte Carr[1], Oliver Smith[1], Sheila Bird[2]
[1] NHS Digital, [2] MRC Biostatistics Unit, Cambridge

High quality statistics, analysis and advice help Britain make better decisions. The challenge for producers of statistics is to get the right data and information at the right time into the hands of those who use them to make a difference and improve care and health outcomes During the H1N1 influenza pandemic in 2009, the UK's Scientific Advisory Group in Emergencies could not access timely information on the deaths of infants (0-4 years) and young children (5-14 years) in England and Wales because of the late registration of inquest deaths. Accordingly, in 2010, evidence by the Royal Statistical Society to the Inquiry into Scientific Advice in Emergencies by the House of Commons Science and Technology Committee highlighted the need for England and Wales to achieve the timely registration of all deaths. In 2017 Professor Sheila Bird, working with NHS Digital and Office for National Statistics, undertook analyses to investigate if information on informal data of death on NHS Digital's Personal Demographics Service (PDS-DOD) might be used as a key source for more timely data than afforded by the date of death as ultimately registered by Office for National Statistics. Initial analysis explored the comparative timeliness and accuracy of these respective data sources and concluded that, for deaths in 2011-14, PDS-DOD was insufficiently available and insufficiently accurate. However recommendations of additional analyses were made to understand why. This talk describes the results of these most recent analyses, for deaths in 2011-16. By linking two rich national person level data sources, we consider the influence that the primary use and business processes associated with each has in relation to timeliness and accuracy of these death-dates reported within. Does data entry error or demographic characteristics of the departed influence timeliness? Finally, we reflect on how the results inform onward use for monitoring and research both now and in the future.

Rapid Fire Talks 3

Tuesday 3 September 3.40pm

Optimisation using emulation in disease modelling. How to use potential improvements to identify new design points.

Daria Semochkina
University of Southampton

The primary interest of this project is in optimising computationally expensive (and possibly competing) noisy objective functions in disease modelling and decision-making. A decision maker potentially could be interested in minimising deaths and minimising the cost of an intervention. Those would most likely be competing goals.The simplest emulation-based optimisation recipe goes like this: we build a sampling plan, we then calculate the responses in these points and fit an emulator to this data, then we can locate the input parameters that are as close to the true minimum of the function as we can make it. However, the optimisation task is not complete until we validate this function value against the true, expensive function itself. We then can, and indeed should retain more of our computational budget so as to allow this search and update process to be repeated many times, adding multiple so-called infill points. When performing a search for a new candidate, we wish to position the next infill point at the value which will lead to an improvement on our best-observed value so far. One possibility is to calculate a so-called expected improvement at a finite set of points and select the best one. This, however, implies that our current minimum observation is the correct representation of the function's value at that point. If there is noise present, using only the noisy observations might be risky, since the noise may introduce errors in the ranking of the observations. In this talk, I will discuss different approaches to emulation-based optimisation that are designed to overcome this drawback in applications to optimal decision making.

Rapid Fire Talks 3

Tuesday 3 September 3.40pm

Analysing time-to-event data with recurrent events

Christiana Kartsonaki
University of Oxford

We consider individuals which are followed up over time and some event of interest is observed. The event may occur more than once during follow-up and observation ends either at censoring or death. Such data frequently arise in medical studies, but often only time to the first event is considered as the outcome of interest. For example, in a study of cancer patients, patients may have more than one recurrences, or in a study of risk factors for pancreatitis individuals may have several hospitalisations for pancreatitis. We consider various models for assessing the associations of explanatory variables with recurrent events, including exponential based, Weibull based and semiparametric models. The resulting estimates and their precision is compared and the different models are illustrated using a dataset on colorectal cancer recurrences.

Rapid Fire Talks 4

Tuesday 3 September 3.40pm

Research and Application of Data Governance Architecture Based on Activity Theory

Huaihai Hui[1], Des McLernon[2], David Allen[2]
[1] *Chinese Academy of Sciences /University of Leeds,* [2] *University of Leeds*

With the rapid increase in the amount of data generated by enterprises and the diversification of data sources, the complexity of the utilization of such massive and multi-source heterogeneous data is greatly enhanced. Therefore, the issue of how to carry out data governance has become a very important topic. In this research, Activity Theory（AT）is proposed as a methodological and analytical framework to study Data Governance Architecture in the Big Data Platform. First, the study defines the data, and then designs the data classification and data model construction methods. Data can be called the Subject in AT. Second, this study proposes how to develop data management policies, how to design data architectures, and how to set data standards and manage data quality, which is the Rule in AT. Third, the study designed a data management organization consisting of decision makers and people directly related to data, which is called a Community in AT. Fourth, this paper allows the data owner, data supplier, data manager, and data user to be in the Division of Labor of AT. Fifth, this study developed the Tools mentioned in AT, including data acquisition tools, data cleaning tools, data integration tools, data search tools, data analysis tools, and data visualization tools. Sixth, this study presents the results of a data governance architecture that includes data assets, data products, data services, and data distribution, which are defined as Objects in AT. Finally, this study runs this data governance architecture as a complete system on the big data platform of an eighth-largest insurance company in the world.

Rapid Fire Talks 4

Tuesday 3 September 3.40pm

Analysis of clickstream data

Ryan Jessop
Clicksco

Online user browsing generates vast quantities of typically unexploited data. Investigating this data and uncovering the valuable information it contains can be of substantial value to online businesses, and statistics plays a key role in this process. The data takes the form of an anonymous digital footprint associated with each unique visitor, resulting in 10^6 unique profiles across 10^7 individual page visits on a daily basis. Exploring, cleaning and transforming data of this scale and high dimensionality (2TB+ of memory) is particularly challenging, and requires cluster computing. We consider the problem of predicting customer purchases (known as conversions), from the customer's journey or clickstream, which is the sequence of pages seen during a single visit to a website. We consider each page as a discrete state with probabilities of transitions between the pages, providing the basis for a simple Markov model. Further, Hidden Markov models (HMMs) are applied to relate the observed clickstream to a sequence of hidden states, uncovering meta-states of user activity. We can also apply conventional logistic regression to model conversions in terms of summaries of the profile's browsing behaviour and incorporate both into a set of tools to solve a wide range of conversion types where we can directly compare the predictive capability of each model. In real-time, predicting profiles that are likely to follow similar behaviour patterns to known conversions, will have a critical impact on targeted advertising. We illustrate these analyses with results from real data collected by an Audience Management Platform (AMP) - Carbon.

Rapid Fire Talks 4

Tuesday 3 September 3.40pm

Remote fault detection: identify faulty refrigeration units using signal processing and machine learning on smart metering data

Phuong Pham, Laura Shemilt, Kaelon Lloyd, Terry Phipps, Sean Stephenson
Centrica

Background: There is a wide variety of household appliances which are operated using electricity and these electrical appliances are prone to failure, malfunction or degradation in operation efficiency to different degrees. Appliance faults or malfunction can lead to risks of power loss or fire. Therefore, early faults detection is important to avoid any consequences. Fault detection requires a good understanding of appliance consumption's profile. While using individual monitoring device for each appliance can be expensive, using Non-Intrusive Load Monitoring techniques on the large volume of smart metering data can help to disaggregate energy consumption for appliance from total household consumption at lower cost.

Objective: We have developed a prototype method to detect faulty fridge freezers using a residential smart meter dataset.

Methods: The input data, the household power signal, is sampled at low frequency (10s interval) and no extra sensors were installed. In the first stage, fridge's power signal is isolated from total household level. Due to the nature of cyclical patterns in fridge's power profile, a mixture of signal processing technique and machine learning, such as spectral analysis and grid search, can be used. In the next stage, the extracted fridge cycle is examined against historic behaviour to detect any outliers from the normal state. Outlying points were detected using several methods including forecasting (arima), global outlier detection (modified z-score) and local outlier detection (local outlier factor, density-based spatial clustering of applications with noise and relative density outlier score). Outliers, which relate to faults in the fridge, can be used to detect breakdowns remotely. The algorithm was developed in R using OOP paradigm to allow fully production.

Conclusion: Fridge’s cycle was derived from total household consumption and it was used in a model combines of different outlier detection techniques. The output from the model enable engineers to identify faulty appliances remotely.

Rapid Fire Talks 4

Tuesday 3 September 3.40pm

Data Science Education, Skills and Industry in Europe

Berthold Lausen[1], Alexander Partner[1], Stephen Lee[2], Henrik Nordmark[3], Mahdi Salhi[3], Christopher Saker[12]
[1] Department of Mathematical Sciences, University of Essex, [2] Mathematics in Education and Industry (MEI), [3] Profusion – a data science and marketing services company

Classification societies focussed on methodologies as for example classification/supervised learning, clustering/unsupervised learning and multidimensional scaling defining the mathematical foundations of the emerging discipline data science, since the Classification Society was founded in London 1964. A special issue of the Journal of Data Science and Analytics (Flach et al., 2018) has papers devoted to current issues in Data Science viewed from a European perspective which were in the European Data Science Conference (EDSC), an invitation only event organised by the late Professor Sabine Krolak-Schwerdt and her team in November 2016 in Luxembourg as the inaugural conference of the European Association for Data Science (EuADS). In this context we discuss challenges and needs for data science education and skills.The landscape for school mathematics in the United Kingdom is very different now compared to even just five years ago. During the period of recent changes in the UK school system we have seen the emergence of undergraduate (BSc) and postgraduate taught (MSc) qualifications in Data Science. For example in 2014 the Department of Mathematical Sciences and the School of Computer Science and Electronic Engineering at the University of Essex have introduced a BSc in Data Science and Analytics and an MSc in Data Science. The curriculum of these courses covers compulsory modules from computer science and mathematical sciences, introducing students to a range of mathematics and statistical topics as well as computing skills such as artificial intelligence, deep learning, information retrieval, programming, and text analytics. In this talk we will also review curricula of data science related university degrees, consider how their content matches industry expectations and discuss plans for further developments.Flach, P, Spiliopoulou, M, Allegrezza, S, Böhmer, M, Hess, B, Lausen, B (2018), Introduction to the special issue on Data Science in Europe, International Journal of Data Science and Analytics 6:163–165.

Rapid Fire Talks 4

Tuesday 3 September 3.40pm

Adjusting reviewer scores for a fairer assessment via multi-faceted Rasch modelling

Caterina Constantinescu
The Data Lab, University of Edinburgh

Selecting submissions for a conference can be viewed as a measurement problem: in principle, organisers aim to accept the ‘best' submissions, but the precise manner in which this is achieved can vary considerably. It is also common to involve multiple reviewers in the process, and it may not always be the case that all reviewers manage to rate all submissions. Hence, there is a chance that some particularly harsh reviewers may rate the same submission (and put it at a disadvantage), or some more lenient reviewers may happen to rate the same submission and propel it higher in the ranking. A solution to this issue is offered by multi-faceted Rasch models, which view submission scores as a function of not just the quality of the submission in itself, but also reviewer severity. This allows to adjust submission scores accordingly, providing a fairer measurement process. Conveniently, the R package `TAM` allows to estimate this type of model. In this talk, I will walk you through an example of how `TAM` was used on data collected as part of the review process for a data science conference in Scotland.

Rapid Fire Talks 4

Tuesday 3 September 3.40pm

The Use and Interpretation of Statistics in Medical Research: An Evaluation of Medical Students' Attitudes

Alaa Althubaiti
King Saud bin Abdulaziz University for Health Sciences

Objective: To evaluate the medical students' attitudes toward statistics use and interpretation in medical research and to assess sex, age and year of course differences in attitudes toward statistics.

Methods: The Attitudes Toward Statistics in Medical Research (ATSMR) survey for health students was administered. The survey assessed students' attitudes by responding to 32 items using a 7-point scale.

Results: 327 medical students participated in the survey with a 54.5% response rate. Although students appeared to appreciate the value of statistics in their professional career, they had negative to neutral attitudes about their feelings towards statistics, their own intellectual knowledge and skills in statistics, and the difficulty of the subject. Students aged 23 years and above perceived statistics to be more difficult than younger students, and they showed less positive feelings towards statistics. Male students showed more positive feelings toward statistics than females.

Conclusion: It is important for medical educators to allocate the adequate mechanisms and design targeted education in statistics courses that stimulate medical students' interest in the field. Innovative ways to deliver the course contents should be explored to improve student' learning experince.

Rapid Fire Talks 4

Tuesday 3 September 3.40pm

Is there a curse of Aaron Ramsey?

Anthony Masters
Nationwide Building Society

Various newspapers and websites have claimed that there is a 'curse of Aaron Ramsey'.The suggestion is that goals by the Arsenal & Welsh footballer are usually followed by the deaths of famous people.The intention is to test this theory.I collate two lists of notable deaths in the UK, from the Daily Mirror and the BBC. This is compared against game data on Aaron Ramsey, from Soccer Base. Date ranges from a single day (same day) to four days (on the same day, to within three days) are considered.Based on one football season (August 2017 to May 2018), there does not appear to be such an effect. Ramsey scored eight goals in this season, only 2 were followed by notable deaths (from the Daily Mirror list) within one day. This is a lower rate than all two-day periods over that season (45%). Looking at three days after a Ramsey goal, six goals were followed by a notable death, only slightly higher the overall rate in that season (69%).Further seasons will be analysed and included in the presentation.

Rapid Fire Talks 5

Tuesday 3 September 3.40pm

STATISTICAL ANALYSIS OF GENETIC RELATIONSHIP OF NIGERIAN AND KENYAN SHEEP POPULATIONS

Asugha KESTER UGOCHUKWU
FEDERAL UNIVERSITY OF AGRICULTURE, ABEOKUTA, NIGERIA

This study evaluates, the genetic relationship of Nigerian and Kenyan-sheep populations using Ovine-50k SNP-chip. Ten samples each were collected from four indigenous Nigerian (Balami and Uda) and Kenyan (Red Maasai and Dorper) sheep breeds and genotyped using the GeneSeek Genomic Profiler Indice HD Bead chip Quality control was carried out using the Gen Abel package implemented in R software based on the following criteria; MAF 1%, SNP Call Rate 90%, Per Individual Call Rate 90%, HWE, IBD ≥ 90%. The descriptive marker and per id summary function of Gen Abel was used for observed Heterozygosity and expected heterozygosity. The Analysis of Molecular Variance (AMOVA) was also carried out on the different markers. Results from this study indicated that, a total of 47 SNPs on the X chromosome were removed for violating quality control. The remaining 818 SNPs on the X chromosome all passed the 90% threshold of per individual calling rate. Also, a total of 835 SNPs on the autosomal chromosome were removed for violating quality control. The remaining 33,176 SNPs on Autosomal chromosome all passed the 90% threshold of per individual calling rate. The sheep population in two major clades and two subclades corresponding to the sampled population indicating each population holds their distinct genetic background with no relationship within and among population, and infers that the sheep breeds from both countries exhibit different ancestral origin. A 70.13% variation within country, 42.56% variation observed among individuals within populations compare to 60.24% variation within population and 42.23% within individuals.The genetic differentiation for autosomes showed the highest differentiation of 0.113 was observed between Red Maasai and WAD while the lowest differentiation was observed between Red Maasai and Dorper with 0.062 compared to that of sex chromosome with a record of 0.144 observed between Red Maasai and WAD while the lowest differentiation was observed between Red Maasai and Dorper with 0.066.

Rapid Fire Talks 5

Tuesday 3 September 3.40pm

THE IMPACT OF AGRICULTURAL PRODUCTIVITY ON ECONOMIC GROWTH IN NIGERIA

SAHEED ABIDEMI Agboluaje
The Polytechnic, Ibadan, Nigeria

The role of agricultural products in the overall development of a nation cannot be quantified. It is not only seen as a key to poverty reduction and vehicle for promoting equity, fairness and social justice but also helps to supply the essential economic ingredients which are necessary condition for sustainable economic growth. The main objective of this work is to carry out an empirical investigation on the impact of agricultural productivity on economic growth in Nigeria, using annual time series data from 2003 to 2017. The Ordinary Least Square regression method was used to analyze the data. The results revealed that there exists a high, positive cause and relationship between agricultural productivity and economic growth. All the variables both dependent and independent includes, the GDP contribution of the agricultural sector, gross expenditure on agriculture and gross access to bank loans had a positive impact in the Nigeria economy and were also in relation to economic growth using the Karl's Pearson correlation coefficient. The findings show a strong impact on agricultural policy in Nigeria. Therefore, in order to improve the activities of the agricultural sector, government should assist the farmers with special incentives such as provision of adequate funding and infrastructural facilities such as good roads, pipe borne water and electricity to mention but a few.Keywords: Agricultural sector, Economic growth, Gross Domestic Product (GDP), Regression.

Rapid Fire Talks 5

Tuesday 3 September 3.40pm

A performance comparison between empirical variograms in achieving the best valid variogram

Esam Mahdi
Qatar University

Modelling the statistical autocorrelations in spatial data is often achieved through the estimation of the variograms, where the selection of the appropriate valid variogram model (especially for small samples) is crucial to achieving precise spatial prediction results from kriging interpolations. To estimate such a variogram, traditionally, we first compute the empirical variogram (the traditional Matheron or the robust Cressie-Hawkins or the kernel-based nonparametric approaches). In this article, we conduct numerical studies comparing the performance of these empirical variograms. In most situations, the nonparametric empirical variable nearest-neighbor (VNN) showed better performance than its competitors (Matheron, Cressie-Hawkins, and Nadaraya-Watson). The analysis of the spatial groundwater dataset used in this article suggests that the wave (hole-effect) variogram model fitted to the empirical variable nearest-neighbor, VNN, variogram is the most appropriate choice. This selected variogram is used with the ordinary kriging model to produce the predicted pollution map of the nitrate concentrations in groundwater dataset.

Rapid Fire Talks 5

Tuesday 3 September 3.40pm

Holistic approach to defining climate in ecological studies

Michel d. S. Mesquita[1], Jan Ove Bustnes[2], Igor Eulaers[3], Carla Vivacqua[4], André Pinho[4]
[1] Bjerknes Centre for Climate Research, [2] Norwegian Institute for Nature Research, Norway, [3] Aarhus University, Denmark, [4] Federal University of Rio Grande do Norte, Brazil

The study of climate ecology has often focused on indices, such as the North Atlantic Oscillation (NAO), that capture a dynamic feature of the climate system. However, such indices may not represent the multivariate nature of the problem, nor local processes that could be at play. Their focus has also limited understanding of the cause and effect in ecological studies. We have created an alternative approach that: emulates the local climate, reduces complexity, and links variables to mechanisms. Our study focuses on the effect of climate variability in the ecotoxicology of tawny owls. These owls are exposed to environmental pollutants, which affect their overall health and stress levels. These pollutants can be transported through the atmosphere, or through the food web (mice, lemming, insects). This study is unique in that it uses a holistic approach to quantifying 'climate' for ecological studies, i.e.: as a composite atmosphere-land process, influenced by large scale dynamics, that affects the local environment.The data come from the ERA-Interim reanalysis project for the period 1979 to 2017. The monthly means of 11 variables were used, which include atmospheric, snow, and soil processes. We have reduced the dimension through the use of Principal Component Analysis (PC). Three PCs were retained based on screeplot and Kaiser criterion analysis. We have interpreted each PC loading as follows: PC1 represents thermal conditions; PC2 indicates (fair) weather conditions; and PC3 (unfavorable) weather conditions. Seasonal variability is captured through the interaction between PC1 with PC2 and PC3.Each PC was then related back to the large-scale dynamics. In conclusion, our PC-based multivariate index not only captures the NAO index, but it also captures the thermodynamics conditions in the region. It takes into account the full environmental conditions by using atmospheric, soil, and snow variables. Furthermore, it reveals a robust and alternative approach to defining 'climate' in ecology studies.

Rapid Fire Talks 5

Tuesday 3 September 3.40pm

Issues in earthquake modelling.

Zak Varty[1], Jonathan Tawn[2], Peter Atkinson[2], Stijn Bierman[3]
[1] *STOR-i Centre for Doctoral Training,* [2] *Lancaster University,* [3] *Shell Global Solutions*

Earthquakes can be caused both by the motion of tectonic plates and by human activity. Both types are inherently difficult to model and predict because the physical processes that cause them are complex, often disputed and difficult to measure accurately. A statistical approach to seismicity modelling provides a natural and popular way to accommodate these uncertainties. The most common approach in the literature is to use the Epidemic-Type Aftershock Sequence (ETAS) model. This model describes earthquake locations and magnitudes using a marked, self-exciting point process. A parametric or semi-parametric model for the conditional intensity function of this point process can be inferred from an observed earthquake catalogue using either frequentist or Bayesian methods. The choice of parametric forms within this function are typically chosen in line with empirical relationships known as the modified Omori and Guttenberg-Richter laws.We present alternative, more flexible, forms for these laws that allow for improved inference, application to smaller catalogues, and better representation of uncertainty in the final fitted model. Proper inclusion of this uncertainty is of paramount importance when deciding how to act in order to mitigate seismic hazard in the future.

Rapid Fire Talks 5

Tuesday 3 September 3.40pm

Forecasting agricultural product and energy prices: A simulation-based model selection approach

Robert Kunst[1], Adusei Jumah[2]
[1]*Institute for Advanced Studies,* [2] *Central University, Accra, Ghana*

The aim of our contribution is twofold. First, we study whether and to what degree the dynamic interaction between commodity prices and energy prices can be exploited for forecasting. Second, we present informative examples for the simulation-based forecast-model selection procedure. Apart from prediction by competing specifications tobe selected from a small choice set, we also explore forecast combinations constructed from a continuum in the same framework. The simulation-based method explicitly permits letting the forecast model choice depend on the intended time horizon of the forecast. With regard to classical Granger causality, the evidence supports a causal direction from food prices to fuel prices, without feedback and somewhat in contrast to our expectations. This causal link, however, only benefits forecasting accuracy at relatively large sample sizes. Similarly, clear evidence on considerable seasonal patterns cannot be fused to a seasonal time-series model that outperforms non-seasonal rivals. Finally, the simulation experiments generally favor the handling of all price series in first differences.

Rapid Fire Talks 5

Tuesday 3 September 3.40pm

Meta-analysis using simple methods successfully derives the big picture for exemptions of fisheries landing obligation

Mickael Teixeira Alves, David Maxwell, Thomas Catchpole
Cefas

The recent reform of the EU Common Fisheries Policy (CFP) includes a landing obligation of all catches of commercial fish. This policy is however subject to exemptions for “species for which scientific evidence demonstrates high survival rates, taking into account the characteristics of the gear, of the fishing practices and of the ecosystem”. The need for timely evidence led us to synthetize results from past projects that collected data on fish vigour at the point of discarding, to inform measures to improve the survival chances of discarded fish. We analysed fishery and environmental data associated with plaice vigour from ten projects conducted over two years in seas around England and Wales. The analysis included basic plots, correlations, multivariate clustering methods and classification trees. Mixed effects cumulative logit models were fitted to subsets of the data to estimate the direction and strength of association between vigour and covariates by project. The combined results of the approaches helped to draw overall conclusions on the factors influencing the vigour of discarded plaice, highlighting the effects of the fish size and temperature. Despite strong heterogenous results between projects, this work illustrates how a combination of statistical tools and careful interpretation can provide scientific evidence for policy purposes.

Rapid Fire Talks 5

Tuesday 3 September 3.40pm

Generalized Regression Control Chart for Monitoring Crop Production in Nigeria using Asymmetric Distribution

Olatunji Arowolo, Mathew Ekum, Samuel Sogunro
Lagos State Polytechnic, Ikorodu

Recently, Nigeria focused on Agriculture as a way to diversify her economy. Crop production, which is a proxy to measure agricultural output is considered important. So, monitoring and controlling crop production (output) among states in Nigeria including FCT is key. This study aimed at using the generalized regression control chart for this purpose. The specific objectives include, identifying the effect of each contributing variable (predictor variable) to crop production for a specified period; setup generalized regression control chart for crop production in the specified period; examine the points out of control and monitor the cross-sections (states) that could be responsible for the deviation from expected crop production limits.The generalized regression control chart was used rather than the usual conventional control chart. The conventional control chart does not put into consideration factor(s) that affect crop production (response variable). The generalized regression control chart however, considers the factor(s) that affect crop production over a specified period and collapsed the assumption of normality, allowing the usage of other flexibility distributions other than normal distribution. Therefore, Weibull distribution was used as the underlying distribution rather than the normal distribution, because it fitted the crop production data better than normal distribution and it is an asymmetric distribution, which is non negative. The data were collected from National Bureau of Statistics (NBS). The result of data fitness showed that Weibull distribution fit the data better than normal distribution. All the predictor variables, farm-gate price, area planted and fertilizer usage, all have significant effect on crop production. The result of the generalized regression control chart showed that crop production is not in control. This implied that some states production is very low compared to others. It is recommended that the states with low crop production should be encouraged to embrace crop production and government should do a proper monitoring of crop production across the state.

Rapid Fire Talks 6

Tuesday 3 September 3.40pm

Convergence and heterogeneity in global diets

Thai Le
Bournemouth University

Worldwide obesity has almost tripled since 1975. This trend is the consequence of demographic, epidemiological and nutrition changes that have taken place as countries develop and become more globalised. This research examines the global trends in food consumption for main food categories since 1961 using data from the Food and Agriculture Organisation of the United Nations (FAO). Preliminary data review discloses a steady rise in food availability worldwide for the past 50 years, coupled with significant alterations in the structure of the global diets in terms of both macronutrients and individual food aggregates. Evidence shows that national diets are evolving over time and across countries in ways that are both similar yet distinct giving rise to patterns that can be investigated statistically by the means of cluster analysis. This research is the first attempt in food economics literature to present the application of an innovative space-time clustering technique inspired by fuzzy logic and copula functions. Agglomerations of countries characterised by similar dietary trends are identified and the findings are further analysed to uncover both economic and cultural factors that most matter in explaining the pace of dietary change as well as the convergence that is observed globally. The findings will inform the public policy debate on the relationship between diet and obesity and provide evidence to those interested in formulating national policies to promote healthy diets.

Rapid Fire Talks 6

Tuesday 3 September 3.40pm

Population health in a digital age: the use of social media and wellbeing in Wales

Jiao Song, Alisha Davies
Public Health Wales

Technology and health survey is the first nationally representative population survey for Wales to examine usage of different social media platforms with state of health and mental wellbeing in Wales, across population groups. A random probability sampling approach was used to identify a nationally representative household sample of 1,240 individuals aged 16 years and above living in Wales. We used Chi-square and Fisher-Freeman-Halton tests followed by binary logistic regression to investigate the differences on usage across social media platforms and functions across demographic and health groups. We will use estimated marginal means to report the adjusted proportions of individuals who reported each usage. 14.6% of over 16 Welsh population has no internet access at home. 63.5% are super users who are on at least one social media platform on a daily basis or a few times a day. Facebook, YouTube and WhatsApp are the most favourite platforms overall. People like spending time on social networking (Facebook and LinkedIn) and sending messages on WhatsApp. Over 90% of younger generations (16-29 and 30-39) are super users while the proportions for 60-69 and 70+ are 46% and 16.3% respectively. The differences of social media usage across Welsh Index of Multiple Deprivation quintiles are statistically significant at p=.05 but not for gender. Inequalities and differences in social media usage with health status and mental wellbeing are currently being investigated, using self-reported state of health (0-100), Short Warwick-Edinburgh Mental Well-being Scale and long-term health conditions as indexes.

Rapid Fire Talks 6

Tuesday 3 September 3.40pm

Social Class and Language Mastery effect in Proficiency Tests

Kaizo Beltrao[1], Elza Nascimento[2], Monica Mandarino[2], Monica Guerra[2], Renato Cardoso[2]
[1] *EBAPE FGV,* [2] *Cesgranrio Foundation*

Bourdieu and Passeron defended the thesis that the school was the main locus to legitimate and perpetuate class differences. This is reinforced by the multiple proficiency tests used in the literature to monitor public policies, which privilege the use of the formal language as part of the instruments. We propose to use a hierarchical model with ENADE's results on standard Portuguese grades, using as covariates, indicators of students' socioeconomic status and economic independence. These two indicators were extracted from the questionnaire filled by the students previously to the exams using Optimal Scaling and Principal Components.According to the INEP, a department of the Ministry of Education who conducts the exam, the National Assessment of Student Achievement (ENADE) assess undergraduate programs through an "exam administered to students who are finishing these courses in higher education institutions throughout Brazil. The programs are grouped in three representative areas and each year one group is assessed, meaning that each program is assessed every three years." Graduating students of all courses take the exam every third year. All knowledge areas with at least 100 courses and two thousand students, mandatorily participate in the exam. These exams, besides a section on specific knowledge related to the professional area being assessed, poses some questions of general interest. The section of general interest is composed of two short answer questions plus eight multiple choices questions. The short answer questions are graded for content and for the use of standard Portuguese. The Portuguese grade takes into consideration three components: orthography; lexical and syntactical. Though questions contents in different exams are not strictly comparable, we hypothesized that the use of formal Portuguese in the answers is. Confirming Bourdieu and Passeron hypothesis, we find out that socioeconomic status do have an impact on language proficiency, but at the University level it is not so strong as measured at lower educational levels.

Rapid Fire Talks 6

Tuesday 3 September 3.40pm

"Severity to those who confess?": Evidence from China's 6,876 Cases of Intentional Injuries

Mengjie Xu, Fang Wang, Liang Guo
Shandong University, Weihai

"Leniency to those who confess, severity to those who resist" is the oldest and most important tenet of contemporary Chinese criminal justice, which was one of the outstanding features of the Chinese communist revolutionary tradition. The goal is to encourage the actor to repent and confess his/her crimes, and in return, receive reduced punishment to encourage him to turn over a new leaf. However, there is a wide perception that confession usually means the suspect will "rot in jail". This study intends to verify whether it is true that "severity to those who confess" and if so, provide an explanation behind this counter-intuitive phenomenon. We randomly selected trial verdicts of 6,876 intentional injury cases from the databases of China's supreme court over the course 2014-2017. Each verdict was read and manually coded by three law students. We estimated a series of linear regression models with the length of prison sentence as the dependent variable, confession as the key independent variable. After controlling 25 variables that can have an effect on the length of prison sentence, we found that in general, confession is positively associated with the dependent variable (0.16, $p<0.01$). however, if a suspect has turned him/her-self in, then the length of prison sentence can be reduced (-0.20, $p<0.05$). The interactive effect between confession and turning-oneself-in is not statistically significant. That is to say, confession tends to receive an increase in punishment. Many confessions may be police-induced. The finding reflects the poor investigatory capacity of judicial organs and the poor protection of suspects' right to remain silent. Our finding indicates that there is a pressing need for the authority to keep the mandatory electronic recording of interrogations and to protect the rights of criminal suspects and defendants against the possibility of judicial abuse of power.

Rapid Fire Talks 6

Tuesday 3 September 3.40pm

Does Job Insecurity Increase the Likelihood of Getting Married? Evidence from 2,123 Chinese Adults

Chaoying Fu, Xiaoli Xing, Mengjie Xu, Liang Guo
Shandong University, Weihai

Prior studies in the US and the UK find that the decline of stable jobs and the rise in casual employment mean that men and women are now less likely to tie the knot, stay married and have children within wedlock. Our study investigates the link between job insecurity and people's marital status (single, married, and divorced) using a sample of 2,123 adults in China. We estimated a series of multinomial logistic regression models with marital status as the dependent variable and job insecurity as the key independent variable. Job insecurity is measured by a dummy variable (1 for a person has signed a fixed-term employment contract, and 0 otherwise). We also control for the effects of workfare and job satisfaction as well as demographical variables. The results show that job insecurity leads to a higher probability of getting married. That is, marriage seems to be a mean for the Chinese people to cope with the destabilizing effects of insecure jobs. The reason why our finding is different from that in the US and the UK may lie in the cultural difference between East and West. In a collectivist country like China, people tend to trust their potential partners and believe that the emotional and psychological commitment of marriage will serve as the solid foundation for the couple to get their share of weal and woe in life. In addition, we estimated the moderating effect of people's education level. The effect of job insecurity on getting married tends to be smaller among university-educated individuals than those who have completed only secondary studies. Likewise, workfare also tends to reduce the impact of job insecurity on getting married. These two moderators confirm that when living in an unstable job situation, marriage can help individuals who have less "resources" to deal with challenges together.

Rapid Fire Talks 6

Tuesday 3 September 3.40pm

The Dark Side of Community-level Social Capital: Does Civic Participation Hurt Mental Health in China?

Xiaoming Lin, Liang Guo, Chunyi Chen
Shandong University at Weihai

Social capital is a crucial determinant of mental health issues at the individual level, the aggregate level, and their interactions. This study focused on civic participation, a type of structural and horizontal social capital, to examine whether it has an impact on self-rated mental health in rural and urban China. Followed the nested modelling strategy, we estimated a series of two-level random-coefficient linear regressions to explain the variations in mental health status, both at the individual level and at the community level by taking the hierarchical structure into account. Using data from our nationally representative sample includes 10,968 Chinese respondents from 130 county-level communities, we found that the civic participation in 13 activities was positively associated with the mental health at the individual level in urban areas, whereas the impact was not significant in rural areas. That is probably that higher civic participation in urban areas could be manifested through progressive coaching mechanism or knowledge dissemination process to fellow individuals, thereby leading to enhanced self-esteem among individuals. On the contrary, at the community level, enhanced civic participation was found to give rise to reduced mental health in urban areas. An explanation for this could be based on the network resources approach, suggesting that mediated by the subjective social status, a higher level of community civic participation could create extra responsibility or overwhelming burden on an individual. Especially in China, fear of losing “face” has been considered as a key deterrent in an individual’s access to social capital and as a source of psychological stress. The results provide an extended lens to analyze the predictors of mental health in China and emphasize the importance to set relevant policy mechanisms to strengthen social networks.

Rapid Fire Talks 6

Tuesday 3 September 3.40pm

Rethinking Reliability of Psycho-political Indicators in Women Elites for Multi-country Surveys

Rachel Gregory
University College Cork

Methodological issues of surveying elites are well documented, including low response rates and response bias. Compounding these problems is the reliability of responses for women in measures, such as political knowledge and political interest. Samples from voters suggest that under certain conditions respondents construct a self-presentation fitting of sex-role stereotypes. Not only has this not been formally tested in elite populations, but the implications of sex-specific response bias in elites could alter policy decisions when attempting to form policy to overcome discrepancies in democratic representation. Using data from the 2014 PartiRep MP Survey, this research seeks to analyse the reliability of women's responses for direct measures of political ambition in comparison to signalling behaviours of ambition by creating an indexed variable of psycho-political behaviours. A comparison between men and women both within and across fourteen European countries finds that while a direct measure of ambition is valid for male respondents, a more reliable measure for female respondents develops from indexed behaviours, suggesting higher response bias for women elites in comparison to male elites. At the same time, it raises the question of possible overestimation of psycho-political indicators from male respondents. Although ambition can be understood as only one psycho-political indicator impacting women's representation in democratic institutions, this finding questions the reliability of other measures requiring explicit responses from women elites. Variation in this finding between countries cautions the use of statistical information in making sweeping policy change on a larger scale.

Rapid Fire Talks 7

Tuesday 3 September 3.40pm

Election prediction using financial models

John Fry[1], Matt Burke[2]
[1] *Manchester Metropolitan University,* [2] *University of East Anglia, , , , , , , ,*

We discuss approaches to use financial techniques to try and predict the results of recent US and UK elections. A combination of different options pricing and econophysics methods are used. A simple adjustment to the method of Taleb is introduced. However, this mainly reproduces results obtained using raw polling numbers. An adjustment based on the JLS model accounts for the possibility that polls may systematically misprice the probability of certain outcomes.

Rapid Fire Talks 7

Tuesday 3 September 3.40pm

FORECAST ERROR: HOW TO PREDICT AN ELECTION: PART 1: POLLS

Timothy Martyn Hill
Barclays Corporate

The "Forecast Error" series of articles in Significance started examining election predictors in 2015. Each article considered many predictors, but each article covered just one election. In 2018 we started a new chapter in the "Forecast Error" series where we examine an individual class of predictor more closely across many elections. This presentation will cover possibly the most prominent: opinion polls.Opinion polls are not intended as predictors but they are frequently used as such. We note the global accuracy work of others, then we focus on the UK by detailing the evolution of British polls since 1937. We examine the final poll accuracy, and the accuracy per month before the election, for British polls since 1945. We then list some investigative tools, consider future developments, and draw conclusions.

SEE ALSO* https://www.significancemagazine.com/politics/595-forecast-error-how-to-predict-an-election* https://www.significancemagazine.com/files/Forecast-error-polls.pdf

Rapid Fire Talks 7

Tuesday 3 September 3.40pm

Admin data and public perception/views of how we use it.

Neil Henderson
NISRA

This rapid fire talk will briefly explore public perception of Government-held administrative data and its use. This talk is provided in the context of administrative data that the Census Office of Northern Ireland holds and explores the Census Office engagement with the wider public. The talk will discuss key concerns raised by the public as identified through previous public engagement through the use of studies and focus groups. Issues like consent, data protection, individual privacy, ethics and using data appropriately were mentioned. Public support for the use of admin data is intrinsically linked to trust; trust in organisations, trust in data protection systems and the perceived benefit to society. Clearly our role must be to satisfy public demands, to improve levels of trust and to clearly and simply explain the benefits to society of Government use of administrative data. Minimising communication about the use of admin data would be counter-productive and would lead people to suspect the government of trying to hide something. This needs to be balanced with the concern that the subject matter is so complicated that it would be difficult to achieve a level of understanding amongst the general public that would prevent people from worrying unduly about their privacy and the security of their personal information.

Rapid Fire Talks 7

Tuesday 3 September 3.40pm

Connected Open Government Statistics

Bill Roberts
Swirrl IT Limited

The strategy of the ONS is "Better statistics, better decisions": that objective of using statistical data more effectively is common across a range of public sector organisations. However, applying statistical data to important questions of public policy or allocation of resources usually requires data from a range of sources. Users report that it is often hard to find the right selection of available data to feed into analysis. Different data providers are inconsistent in their approach, making combination or comparison of data difficult.Swirrl is working with the ONS and the Government Statistical Service in the 'COGS' project ('Connected Open Government Statistics') to develop practical solutions to these problems.At the heart of our approach is delivering data so that it works well in software, combined with the use of standards to enable interoperability. The challenge is choosing or designing standards and then getting everyone to follow them. The project has reviewed and prototyped standard approaches for structuring and exchanging data, building on the standards of the web.We now have a set of processes and software tools to support them, that government data publishers can follow to make their statistical data accessible in a richer way. Interchange processes are based around CSV files following the 'Tidy Data' structure. Re-usable codelists and variable definitions are prepared and made easily findable. Behind the scenes, the standards of Linked Data are used to store and organise the data in an interoperable way. The project is liaising with Reproducible Analytical Pipelines teams at various departments, to integrate this new dissemination approach with modern techniques for preparation of statistical releases.This talk will explain the details of the approach and report on early experiences of implementation.

Rapid Fire Talks 7

Tuesday 3 September 3.40pm

Bad evidence encourages bad policy

Paul Marchant
Leeds Beckett University

Policy that is based on evidence is surely the rational choice. However if the evidence put forward is flawed, more harm than good may result from a policy. It may be hard to get an unjustified policy changed, because of the illusion of an evidential basis. In addition, personal investment by those having encouraged the policy may hinder its change. High standards of research are essential to ensure only trustworthy evidence is used when deciding policy. Policy makers must be attuned to the need for separating wheat from chaff.In this presentation, the flaws in some studies used to encourage expenditure will be identified.Some particular errors include:Definitively ascribing cause from correlation in an observational study with a weak design. Performing cosmetic analyses on grouped and averaged data, rather than individual cases, achieves an implausibly high coefficient of determination (R2). The study is cited elsewhere in a bid to acquire huge government funding.In another report, that of an RCT, the control arm is not used, because it is ‘not statistically significant’, when generating the size of the intervention effect. However, clues in the report enable an estimate the effect of the omission. Other statistical misdemeanours are also committed. The work again encourages public money to be spent on a false promise.Existing best practice would go some way to reduce risks of believing bad evidence through having: 1) a transparent public protocol written before embarking on the research and 2) open research data to check work and to enable better analyses.A way of protecting against continuing with implementing flawed policies would be to have, a pre-defined, scientifically rigorous plan to check the effects of the roll out of any policy.

Rapid Fire Talks 7

Tuesday 3 September 3.40pm

Measuring the Economy - An ONS online book

Georgia Tasker-Davies, Jonathan Athow, Ed Palmer
ONS

Measuring the Economy The Office for National Statistics is creating an online book ‘Measuring the Economy’ for use by universities to support undergraduate and postgraduate economic statistics teaching. The book aims to convey the importance of best-practice, real-world economic measurement, and as such to support academics and students in understanding the issues surrounding the measurement of the modern economy. Chapters are being written by leading experts in their subject area, with the aim being to bring their knowledge and expertise together in a single output. Our aim is for the chapters to be standalone, so that they can be used a la carte while, also maintaining a common and coherent approach. We are looking to develop teaching resources for each chapter- case studies, exercises and test questions for example, once the chapter has been published in Beta. As soon as these additional resources have been produced we will make them available, alongside the chapter, for use and comment. Our aim is to publish the chapters, in Beta, between May and December 2019. During this period, and up to March 2020, we are inviting feedback and review from Stakeholders. We wish to collaborate with these stakeholders to shape the project and ensure the final product is beneficial to the University community. We will engage with this community regularly throughout the project, gaining continual feedback and review of each chapter as it is published. Publication of the online book, in alpha, is anticipated to be September 2020. The ‘Measuring the Economy’ intention is also to publish a regularly update hard copy thereafter. Chapters:IntroductionMeasuring inflationNational Income and ExpenditureSectoral AccountsSupply-use FrameworkThe Labour MarketProductivity and GrowthInequalitySub-National StatisticsTrade and Balance of PaymentsThe Financial SectorThe Environment – including Natural CapitalHard to Measure SectorInnovation StatisticsPublic Finance

Rapid Fire Talks 7

Tuesday 3 September 3.40pm

Faster Indicators of UK Economic Activity by using over a hundred million VAT returns

Luke Shaw, Louisa Nolan, Jeremy Rowe, Andrew Sutton, Daniel Ollerenshaw, Stephen Campbell, Ioannis Kaloskampis
Office for National Statistics (ONS)

There is a current great appetite for faster information on UK economic activity. Indeed, the Independent Review of Economic Statistics (Bean, 2016) stated that “the longer a decision-maker has to wait for the statistics, the less useful are they likely to be”. The faster indicators of UK economic activity project is one of the Office for National Statistics’ innovative responses to this demand. The project identifies close-to-real-time large administrative and alternative data-sets that are related to important economic concepts. From these data-sources, we develop a set of timely indicators that allow early identification of potential large economic changes. Data-sets included thus far are: HM Revenue and Customs Value Added Tax (VAT) returns, ship tracking data from automated identification systems for UK waters, and road traffic sensor data for England. Here we focus on indicators built from VAT returns data. We have constructed monthly and quarterly diffusion indices built using turnover and expenditure data from VAT returns, and several indicators based on VAT reporting behaviour. We discuss the methodology behind the indices and present results-to-date. We caution that care should be used in interpreting these indicators, and they are supplementary to, not a proxy for gross domestic product. However, the suite of indicators shows promise in identifying large changes to economic activity. We find that, at the time, the VAT quarter-on-quarter turnover diffusion index would have identified the first quarter of the recession which begun in 2008 five months before it was seen in official estimates. Since April 2019 we have been publishing these indicators monthly, within a month of the end of the period of interest. This is one month in advance of official GDP estimates.

4.1 Contributed - Medical: Survival Analysis

Wednesday 4 September 9am

Landmark analyses of survival benefit associated with statin prescription

Ilyas Bakbergenuly, Elena Kulinskaya, Lisanne Gitsels
University of East Anglia

Objective: Statins are prescribed for primary and secondary prevention of cardiovascular disease, however the threshold of cardiac risk at which to prescribe statins is still controversial, especially at older ages where everyone would be eligible. The research objective was to dynamically predict the survival benefits associated with statin therapy over the course of 25 years in patients residential in England or Wales.

Methods: Primary care records from The Health Improvement Network (THIN) were used. The cohort included 110,243 patients who turned 60 between 1990 and 2000, were neither diagnosed with cardiovascular disease nor prescribed statins. The cohort was followed up until January 2017, where the medical history was updated every half a year. Landmark analyses were carried out by fitting adjusted Cox's proportional hazards regressions of the hazard of all-cause mortality associated with current statin prescription at each landmark from age 60 to 85 (51 time points).

Results: The initial results show that statin therapy is associated with increasing survival benefits by older ages. By age 65, statin therapy became significantly associated with survival (age 60 HR=0.88 (0.67-1.18) and age 65 HR=0.86 (0.78-0.94)). This survival benefit remained approximately the same for the next ten years (age 70 HR=0.85 (0.79-0.91) and age 75 HR=0.81 (0.74-0.89)), after which greater benefit was gained (age 80 HR=0.72 (0.62-0.84) and age 85 HR=0.57 (0.35-0.91)). The survival benefit of statin therapy significantly differed by the birth cohort but not by sex or cardiac risk. The 1936-40 cohort showed larger survival benefits of statin, probably due to development of more efficient drugs. Furthermore, current prescription was a stronger predictor than cumulative proportion of statins over time and age at first prescription.

Conclusions: After adjustment for cardiac risk and related medical history, it appears that statin therapy is especially beneficial at older ages. Clinicians may want to use this new information when managing cardiac risk of older patients.

4.1 Contributed - Medical: Survival Analysis

Wednesday 4 September 9am

Identifying biomarkers to predict pancreatic cancer

Christiana Kartsonaki
University of Oxford

Pancreatic cancer has the worst overall prognosis of all cancers, with a 5-year survival less than 5%. Several metabolic and lifestyle factors are associated with pancreatic cancer risk, but there is need to identify biomarkers that may help with risk prediction and early diagnosis of pancreatic cancer. We aimed to identify protein biomarkers in blood which are associated with being diagnosed with pancreatic cancer up to several years later.We designed a case-subcohort study within the China Kadoorie Biobank, a prospective cohort study of over 500,000 Chinese adults with blood samples collected at baseline, to examine the associations between circulating proteins and the risk of developing pancreatic cancer. We used the OLINK immuno-oncology proteomics assay to quantify 92 biomarkers in 700 pancreatic cancer cases that accumulated over about 8 years of follow-up and a randomly sampled subcohort of 700 individuals. We used Cox proportional hazards models with the Prentice pseudo-partial likelihood to assess the associations between proteins and pancreatic cancer risk. Time in study was used as the timescale. We examined the shape of the associations using splines and by splitting protein values into groups. We used forward selection and the method by Cox and Battey (2017) to identify sets of proteins that predict pancreatic cancer risk.We identified several proteins that are associated with pancreatic cancer risk. Sets of proteins identified from different variable selection approaches largely overlapped. Sensitivity analysis with different weighting methods, using age as the timescale or adjusting for other risk factors did not substantially change the conclusions. Including the identified markers into models with other risk factors yielded a modest improvement in the discriminatory ability of the model.We identified biomarkers which may help predict risk of pancreatic cancer up to several years after measurement and which may contribute to understanding its aetiology.

4.1 Contributed - Medical: Survival Analysis

Wednesday 4 September 9am

Is the Restricted Mean Survival Time Approach an Alternative to the Time-Dependent Cox Model When Hazards are Non-Proportional?

Bee-Choo Tai[1], Zhaojin Chen[1], Joseph Wee[2]
[1] *National University of Singapore,* [2] *National Cancer Centre*

Objective: We evaluate the performance of the restricted mean survival time (RMST) approach as an alternative to the Cox time-varying covariate (TVC) model for quantifying the treatment effect when there is non-PH.

Methods: To simulate non-PH, survival times were generated using Gompertz distribution with shape parameter α = 0.1, scale parameter γ = 0.25, for n = 200 and 500 with 1000 replications. The censoring distribution was generated from U(0, 6), and treatment B(1, 0.5). The life-expectancy ratio (LER) was used to quantify the effect of treatment. Assuming LER from Gompertz distribution as the gold standard, the RMST estimates from the flexible parametric survival model with 3 knots and 1 df were compared to those of Cox TVC in terms of bias and mean squared error (MSE) for t = 1, 3 and 5 years. The two methods were also compared using data from a randomised clinical trial of patients with nasopharyngeal cancer (SQNP01).

Results: The RMST and Cox TVC estimates were close to the true LER for all t, although the former had slightly larger bias and MSE. When applied to the SQNP01 data, LER of both models increased with time. At 1-year, the estimates based on RMST and Cox TVC were 0.99 (95% CI 0.95 to 1.02) and 1.00 (95% CI 0.96 to 1.04) respectively. Their 3- and 5-year estimates of 1.04 (95% CI 0.96 to 1.13) and 1.04 (95% CI 0.97 to 1.14), and 1.16 (95% CI 1.04 to 1.30) and 1.18 (95% CI 1.05 to 1.32) respectively were also progressively larger.

Conclusion: The RMST approach yielded slightly larger bias and MSE as compared to Cox TVC when the hazards are non-proportional. However, estimate based on RMST may be more appealing to clinicians because of its ease in interpreting the magnitude of survival benefit.

4.2 Contributed - Official and Public Policy: Alternative data sources

Wednesday 4 September 9am

Estimating the impact of automation

Andrea Lacey, Anna Ardanaz-Badia
Office for National Statistics

Automation has become more relevant in recent years due to enhanced technology, including data science and AI, which allows various processes to be automated. This is a matter of both public and policy interest, as the introduction of new technologies will have implications for the UK labour market. We have utilised an OECD study that assesses the probability of automation of the tasks that people undertake within certain occupations, and applied the UK outputs to the Annual Population Survey. Using this, we've produced estimates of the risk of automation for occupations in 2011 and 2017 by various demographics. In addition, we've looked at the skills within an occupation that are at high and low risk of automation, and what skills may be 'transferrable'. This presentation will discuss the methodology used to produce estimates of automation in the UK, and the results of our analysis, including what jobs are at risk of automation, and who is most likely to be employed in these roles. We will also explore regionally where automation may affect.

4.2 Contributed - Official and Public Policy: Alternative data sources

Wednesday 4 September 9am

Young People's Earnings Progression and Geographic Mobility

Bonang Lewis, Thomas Odell
Office for National Statistics

This presentation will be based on the published article 'Young people's earnings progression and geographic mobility'. Innovative new Census linked to administrative data was used to connect longitudinal income data to personal characteristics, allowing more granular and intersectional analysis on how multiple forms of earnings progression differed for young people between 2011/12 and 2015/16. Regression analysis identified characteristics of people more likely to escape low pay. We also investigated patterns of movement between local authorities for young people, and how earnings growth varied depending on starting area and different city-region destinations. This work informed policy thinking around the Department for Work & Pensions Labour Market strategy and provided evidence for the Race Disparity Unit, to target funding based on ethnic disparities affecting education to labour market transitions. The work was a successful feasibility test for this unique admin linked dataset as it provided helpful new insights.

4.2 Contributed - Official and Public Policy: Alternative data sources

Wednesday 4 September 9am

Advancing the methods for administrative and transactional data in official statistics

Hannah Finselbach, Lucy Tinkler
Office for National Statistics

The Office for National Statistics (ONS) is transforming to put administrative and alternative data sources at the core of our statistics. Combining new sources with surveys will allow us to meet the ever-increasing user demand for improved and more detailed statistics. Our history is mainly of indirect usage - in small area estimation, calibration weighting, census quality assurance, and multi-source migration patterns. More recently we've become more ambitious - and successful - in directly replacing (and augmenting) our high-quality survey statistics with admin data. Like many national statistics organisations, we want to exploit the rich data sources available from transactions by businesses and households with government agencies and digital platforms. However, using this data involves addressing a range of statistical challenges, including those described by Hand (2018)*. ONS has established a methodological research programme, to develop a theoretical framework to effectively use and integrate new data sources. We aim to work collaboratively across academia, government and private sector on a variety of projects, to address some of these statistical challenges, including:

- Measuring and communicating uncertainty of integrated or administrative data
- Linking multiple sources for multiple uses and analysis
- Measures the quality of data linkages
- Managing discontinuities within the data
- Creating synthetic data to enable researcher access to confidential data
- Dealing with definitional differences between admin sources and survey data
- Identifying and addressing under or over coverage
- Ensuring confidentiality and privacy of linked dataIdentification and correction of errors in big/streamed data

This talk will outline our achievements to date, our plans, and the challenges that remain to be solved.

*Hand, D., “Statistical challenges of administrative and transaction data.” Journal of the Royal Statistical Society A, 2018

4.3 Contributed - Applications of Statistics: Applications 1

Wednesday 4 September 9am

Supervised Classification of Linear Synchronous Motor Vehicle State in a Smart Factory.

Jill Daly
CIT

The research presented in this thesis examines how to use Vibration and Magnetometer sensor data to identify where on a manufacturing production line a Linear Synchronous Motor vehicle is situated and the type of movement of the vehicle at that position. This paper seeks a solution to this problem by identifying the state of the vehicle on the track from the data observations generated by a data capture experiment. The results from this research will be part of a predictive maintenance pipeline in a smart industry environment.In the absence of a labelled dataset, work was undertaken to label approximately 1.5 million sensor data observations. This labelled dataset was used for iterative model training, validation and feature engineering. Due to the large amount of data available, it was possible to keep a section of data separate, so that it could be used as unseen validation data. This validation data was additional to the 10 fold cross validation used. Model performance was measured using accuracy and specificity because type 2 (false positive) errors were identified as the most costly errors. Multiclass One-v-Other ROC was used as a further measure of model performance.The algorithm adopted was the Supervised Learning Random Forest classifier. The final Random Forest model was trained using data for Vibration and Magnetometer combined, with all speed settings. The features used were rolling aggregation mean and standard deviation, derived from the Z_AXIS, and with FFT applied to those aggregate values. This model provided a training accuracy of 85.6% and a validation accuracy (i.e.: generalisation) of 95.7%.

4.3 Contributed - Applications of Statistics: Applications 1

Wednesday 4 September 9am

Modelling Road Accidents in Edinburgh Using Hidden Markov Models

Valentin Popov[1], Glenna Nightingale[2], Andrew Williams[3], Paul Kelly[4], Ruth Jepson[2]
[1] University of St Andrews, [2] School of Health in Social Sciences, Medical School, Edinburgh, UK, [3] European Centre for Environment and Human Health, University of Exeter Medical School, [4] Physical Activity for Health Research Centre, University of Edinburgh

Empirical study of road traffic collision rates is challenging at small geographies due to the relative rarity of collisions and the need to account for secular and seasonal trends. In this paper we demonstrate the application of Hidden Markov Models (HMMs) to successfully describe road traffic collision time series using data from the city of Edinburgh (STATS19) as a case study. We apply various other time series models to these data and demonstrate the superiority of the HMMs for this particular case. Our model discrimination approach, based on the Akaike Information Criterion indicates that the best performing model is a 2-state HMM with Negative Binomial state-dependent distributions. The model accounts for the diminishing trend observed in the data and, contrary to other competing models, indicates teh presence of seasonality effects. In addition it shows good forecast ability. To date, Hidden Markov models have not been used to model road traffic data from the UK. The application of HMMs to such routinely collected data may be beneficial to natural experiments or evaluations of interventions and policies that seek to impact traffic collision rates.

4.3 Contributed - Applications of Statistics: Applications 1

Wednesday 4 September 9am

Dynamic Spatial Sampling in Semiconductor Manufacturing

Seán McLoone[1], Gian Antonio Susto[2]
[1] *Queen's University Belfast,* [2] *University of Padova*

Semiconductor manufacturing typically involves silicon wafers undergoing hundreds of different steps over several weeks to build up the desired complex nanoscale structures. Metrology is essential for monitoring the performance of these processes to ensure that both spatial variation (intra-wafer) and temporal variations (inter-wafer) do not exceed the demanding tolerances required for modern semiconductor devices. However, metrology is a high cost, time consuming, and non-value added operation, hence, due to commercial pressures, standard practice is to keep metrology to a minimum by adopting both spatial and temporal sampling protocols. Here we consider the problem of optimising spatial dynamic sampling plans such that the number of locations measured on each wafer is minimised while retaining the capacity to: (1) generate an accurate reconstruction of the wafer profile, and; (2) to detect previously unseen process behaviour in a finite time horizon.The solutions we have developed are data driven, and involve the analysis of historical metrology data using Forward Selection Component Analysis (FSCA), a greedy search based unsupervised variable selection technique. Using this algorithm, a number of strategies for generating dynamic sampling strategies are explored, and the trade-off that exists between wafer profile reconstruction accuracy and the ability to detect previously unseen anomalies assessed.Results for both simulated and industrial metrology case studies confirm the efficacy of the proposed FSCA based dynamic sampling methodology, with substantial reductions in the number of measurement sites that need to be measured possible for processes that exhibit significant spatial correlation in inter-wafer variability. In particular, analysis of a metrology dataset from a chemical vapour deposition process used in the manufacture of hard disk drives shows that measurements per wafer could be reduced from 50 to 7 while maintaining wafer profile reconstruction accuracy levels of greater than 99%.

4.4 Contributed - Social Statistics: Populations

Wednesday 4 September 9am

Measurement Error Model to Correct the Inconsistencies in Migration Flow Data for South America

Andrea Aparicio-Castro[1], Arkadiusz Wiśniowski[1], Francisco Rowe[2], Mark Brown[1]
[1] *University of Manchester,* [2] *University of Liverpool*

The intra-regional migrant stocks in South America doubled from 2 in 1990 to 4 million people in 2015. This evidences that a great number of migrants of the region changed their migration strategy towards South America. To gain an understanding of the spatial patterns and trends underpinning the increasing appealing of immigrants to South America, reliable data on migration flows are crucial to measure the evolving migration links between countries of origin and destination.While migration flow data are available from individual countries, they often incomplete and/or incomparable between countries and over time. Data sources differ not only in how they define migrants and their population coverage, but also in their systematic bias, accuracy and measurement methods, such as censuses, population registers and surveys.To address these issues, this paper aims to develop a measurement error model to correct inconsistencies in migration flow data for South America. Multiple migration data tables from a variety of sources are combined in a measurement model to correct and measure true (unobserved) migration flows. The model captures inconsistencies in duration of stay, systematic bias in measurement, population coverage, accuracy and variability of reported flow data from each data source. The resulting outcome is a set of synthetic estimates of migration flows with measures of uncertainty.

4.4 Contributed - Social Statistics: Populations

Wednesday 4 September 9am

Combining health information systems data and probability survey data to monitor health coverage indicators in low-resource settings

Caroline Jeffery[1], Marcello Pagano[2], Baburam Devkota[1], Joseph J Valadez[1]
[1] *Liverpool School of Tropical Medicine,* [2] *Harvard T.H. Chan School of Public Health*

Objectives: Delivering health services in low-resource settings partly depends on accurate Health Information Systems(HIS) data. The quality of these records matters because poor quality leads to poor judgements and outcomes. Unlike probability surveys, which are representative of the population and carry accuracy estimates, HIS do not. They may not even detect gaps in service coverage and leave communities exposed to unnecessary health risks. Using data from Benin, Madagascar and India, we improve informatics by introducing a novel method, the Annealing Technique (AT), for combining HIS and survey data that improves the accuracy of health coverage indicators at the district and sub-district levels, where management improvements are made, and provides an all-important measure of the accuracy of the indicators.

Methods: The AT estimator is a weighted average between the administrative and the survey estimators. The two weights are inversely proportional to their respective variances. We apply AT to 3 datasets, where the survey data were collected using Lot Quality Assurance Sampling: polio vaccination and vitamin A distribution in 19 communes of Benin and 3 districts of Madagascar (2015); and 10 indicators measuring pre-natal, delivery and post-natal care in 25 blocks in Bihar, India (2016).

Results and Conclusion: For Benin and Madagascar, the annealed and survey estimates of the communes or districts differ by no more than 3%, while decreasing the standard error (SE) by 1% to 6%; the HIS and annealed estimates are quite different, with 3 to 29 times larger SE. For Bihar, the annealed and survey estimates of the blocks differ by no more than 10%, while decreasing the SE by 0.05% to 32.4%. The HIS and annealed estimates can differ by up to 84.2%, with 1.4 to 32.3 times larger HIS SE. The AT methodology creates more accurate estimators, together with measured HIS errors, marking a new role for probability sampling and HIS for tracking the public's health.

4.4 Contributed - Social Statistics: Populations

Wednesday 4 September 9am

Modelling and forecasting UK fertility using Bayesian Generalised Additive Models

Joanne Ellison, Ann Berrington, Erengul Dodd, Jonathan J Forster
University of Southampton

Aggregate UK fertility data available at the population level can only give limited information about the patterns of variability of age-specific fertility rates. Survey data provides a rich source of information through fertility histories of individuals and their corresponding characteristics, which can help to gain a better understanding of the underlying variability of fertility rates. By modelling the fertility histories of women surveyed in Wave 1 of the UK Household Longitudinal Study (Understanding Society), we investigate the dependence of birth events on selected variables as well as information derived from the fertility histories themselves. Generalised Additive Models (GAMs) allow the incorporation of covariates and interactions between them as smooth terms, where the precise form of the smoothness is purely data-driven. Fitting Bayesian parity-specific logistic GAMs to the survey data, we learn about the variability of fertility as a function of age, cohort, time since last birth, highest educational qualification and country of birth. Following model selection, we find that educational attainment is important for determining the likelihood of transitions to lower order births, whereas country of birth significantly influences the odds of transitioning to higher order births. Also, the chosen models decrease in complexity as parity increases. There is the potential to combine inferences from this detailed individual-level data with the coarser population-level data for the purposes of forecasting.

4.5 Contributed - Methods & Theory: Methods Showcase

Wednesday 4 September 9am

Large numbers of explanatory variables

Heather Battey[1], David Cox[2]
[1] *Imperial College London,* [2] *Nuffield College, University of Oxford, , , , , , , ,*

In the context of regression with a large number of potential explanatory variables and relatively few study individuals, a key point is emphasised: if there are several models that fit the data essentially equally well, one should aim to specify as many as is feasible. This view is in contraposition to that implicit in the use of variable selection methods, which produce a single model effective for prediction but potentially misleading for scientific understanding. I will discuss a different approach whose aim is essentially a confidence set of models. The talk is based on joint work with David R Cox:Cox, D.R. and Battey, H.S. (2017). Large numbers of explanatory variables, a semi-descriptive analysis. Proc. Nat. Acad. Sci., 114 (32), 8592-8595.Battey, H.S. and Cox, D.R. (2018). Large numbers of explanatory variables: a probabilistic assessment. Proc. R. Soc. Lond. A., 474The work is motivated by genomics. However, the ideas are mathematical in basis and apply more widely.

4.5 Contributed - Methods & Theory: Methods Showcase

Wednesday 4 September 9am

Spaed rankings, and how to assess them --- with application to predicting the Premier League

David Firth
University of Warwick

This is methodology inspired by an attractive --- and now quite commonly seen --- method for summarising a predictive distribution on the possible rankings of N items. An example, predicting end-of-season standings in Premier League football, can be seen at https://twitter.com/DectechSports/status/1099241216221286401. The prediction is presented as a nonnegative NxN matrix with row and column sums all equal to 1 (or 100%) --- a 'doubly stochastic' matrix. This provides a readily understood summary of the predictive distribution. In the Premier League example, each row is a probabilistic forecast of the finishing position for a given team, and each column predicts which teams could ultimately occupy a specific league position. When a doubly stochastic matrix is used in this way we will call it a 'spaed ranking'.(Spaed is the past tense of an old Scottish verb 'spae', meaning predict or foretell. Its use here is motivated by the need to distinguish a spaed ranking from a full predictive distribution over the set of all possible rankings. A spaed ranking can be viewed as the expectation of such a predictive distribution.)We develop the use of a proper scoring rule (such as the familiar Brier score), to compare spaed rankings and also to assess the performance of a spaed ranking in and of itself. The latter assessment is possible because a spaed ranking gives probabilistic forecasts for several outcomes simultaneously. We show how to calibrate any chosen scoring rule through its sampling distribution under the 'canonical' predictive distribution consistent with a spaed ranking. The sampling distribution is analytically intractable, but is amenable to accurate approximation through an intuitive MCMC scheme.The methods are general --- not limited to sports applications! --- but will be illustrated here through application to published spaed rankings in Premier League football.

4.5 Contributed - Methods & Theory: Methods Showcase

Wednesday 4 September 9am

Inference under unequal probability sampling with the Bayesian exponentially tilted empirical likelihood

Andrew Yiu, Robert Goudie, Brian Tom
MRC Biostatistics Unit

In this presentation, we consider estimating a population outcome mean in the presence of unequal probability sampling. We work under two settings. The first assumes a selection mechanism determined by the data collector, but the analyst is provided only with partial design information in the form of sampling probabilities for the selected individuals. The second is an observational setting where the selection mechanism is unknown but is assumed to be independent of the outcome conditional on a set of fully observed covariates. Certain problems in missing data and causal inference, such as estimating the average causal effect of a binary treatment with no unmeasured confounders, can be framed in this perspective. In both cases, semiparametric estimators incorporating inverse probability weighting have attractive large sample properties such as local efficiency and double robustness. However, the reliability of these estimators is less justified theoretically in small sample inference, where the use of an informative prior in a Bayesian approach may be advantageous. We provide an inferential framework that offers the practical benefits of Bayesian statistics, along with the desirable asymptotic guarantees of semiparametric estimators. Central to this method is the exponentially tilted empirical likelihood, which enables Bayesian analysis of moment condition models. We use the fact that many semiparametric estimators can be defined as the solutions to unbiased estimating equations, and formulate a set of moment constraints to calibrate posterior behaviour to the estimator of choice. We prove Bernstein-von Mises theorems for a broad class of moment constraints, establishing large sample equivalence with the frequentist approach. In particular, the frequentist coverage of posterior credible intervals converges to the nominal level, and the posterior concentrates around the limiting value of the estimator. Simulations verify these theoretical results and suggest that our approach outperforms existing Bayesian proposals.

4.6 Contributed - Medical: Identification and Prediction

Wednesday 4 September 9am

Sample size requirements for validating risk prediction models for binary outcomes

Chen Qu, Gareth Ambler, Rumana Omar, Menelaos Pavlou
UCL

Objective: Risk prediction models for health outcomes are used in the clinical management of patients and assessment of institutional performance. Given their importance in health care, it is essential that the predictive performance of a risk model is assessed adequately in an external validation study. However, there has been limited research into the sample size requirements for validation studies and hence we aim to provide guidance to researchers through this work. We propose precision-based methods to calculate the minimum sample size required for an external validation study for binary outcomes.

Methods: The performance of a risk model is often assessed using the calibration slope and c-statistic, hence it is important to estimate these measures with acceptable precision. We derived approximate analytical expressions for the standard error of both estimates in terms of the sample size (N), c-statistic (C) and outcome prevalence (P). The accuracy of the analytical expressions was evaluated in a range of simulation studies varying the values of C, P and N.

Results: The simulation results show that the proposed expressions provide good approximations to the standard errors. That is, the expressions may be used to estimate the precision of the c-statistic and calibration slope for a given sample size, and hence can be used to determine a suitable sample size for a prospective external validation study. For example, our analytical expressions suggest that a sample size of 1800 is required to estimate the c-statistic with a standard error of 0.02, when the outcome prevalence is 10% and the c-statistic is 0.72. With that sample size it is possible to estimate a calibration slope of 1 with a standard error 0.11.

4.6 Contributed - Medical: Identification and Prediction

Wednesday 4 September 9am

The use of period analysis techniques as an alternative approach to estimate post-transplant kidney survival outcomes

Chloe Brown
University of the West of England / NHS Blood and Transplant

Objectives: Organ transplantation is an area that is constantly evolving, with improved survival observed over the last ten years. Nationally, kidney transplant outcomes are monitored and survival estimates, obtained from traditional survival analysis methods, are informative for patients and clinicians. It is therefore important that survival estimates are up-to-date, reflective of current practice and applicable for a recently transplanted cohort. Period analysis is a potential alternative approach which is thought to obtain more up-to-date survival estimates. Here, different approaches for obtaining unadjusted and risk-adjusted five-year kidney graft survival estimates are compared to determine which approach is the most accurate and appropriate.

Methods: Data on adult first deceased donor kidney transplants performed between January 2006 and December 2018 were obtained from the UK Transplant Registry. Data were analysed as shown in Table 1. Five-year graft survival estimates from each approach were compared to the actual observed five-year survival. Cox proportional hazards modelling and a period approach to Cox modelling will be compared in obtaining risk-adjusted survival estimates.

Table 1:

Time period	Method	Analysed as at	Name
2006-2008	Kaplan-Meier	December 2013	Approach 1
2006-2013	Kaplan-Meier	December 2013	Approach 2
2006-2013, window period 2011-2013	Period	December 2013	Approach 3
2011-2013	Kaplan-Meier	December 2018	Actual observed survival

Results:The different approaches produced survival estimates within 1% of the actual observed survival. The unadjusted five-year survival estimates for the three approaches were 85.7% (1), 86.4% (2) and 86.8% (3) respectively, which compared to an actual observed survival of 86.3%.

Conclusion: Although there appears little difference between survival estimates obtained by each approach, period analysis does possess other advantages. By only considering the survival experience in a recent window period, this enables survival estimates of recently transplanted patients to be obtained and earlier detection of changes in survival following changes in practice. Comparison of a standard and period approach to Cox modelling is currently in progress.

4.6 Contributed - Medical: Identification and Prediction

Wednesday 4 September 9am

Analysis of paediatric visual acuity using Bayesian copula models with sinh-arcsinh marginal densities

Mario Cortina Borja[1], Julian Stander[2], Luciana Dalla Valle[2], Charlotte Taglioni[3], Brunero Liseo[4], Angie Wade[1]
[1] *Great Ormond Street Institute of Child Health; University College London,* [2] *School of Computing, Electronics and Mathematics; University of Plymouth,* [3] *Dipartimento di Science Statistiche; Università degli Studi di Padova,* [4] *Dipartimento Metodi e Modelli per l'Economia, il Territorio e la Finanza; Sapienza Università di Roma*

It is important that children who are visually impaired are identified early enough to start corrective treatment. Visual acuity, which quantifies each eye's spatial resolution capacity, was measured in the left and right eyes of over 2700 children, together with each child's age (ranging from 3 to 8 years). The distribution of visual acuity is asymmetric and kurtotic thus needing four parameters to be adequately modelled so the sinh-arcsinh distribution specified the marginals of the bivariate model for measurements from both eyes. A flexible bivariate statistical model was developed using copulas to understand how visual acuity changes with age in each eye separately, and how the dependence between left and right eye visual acuity is related to age. It was found that as age increases, visual acuity improves and the dependence between visual acuity in the left and right eyes becomes stronger, meaning that children's eyes become better and more similar with age. The bivariate statistical model also allowed the identification of children with unusual sight, distinguishing those who are atypical in both eyes when they are considered together from those who are outliers in one or both eyes when they are considered separately, as in current practice. This yields an innovative tool that enables clinicians to recognise children with unusual sight who may otherwise be missed.

4.7 Contributed - Data Science: Misc

Wednesday 4 September 9am

Big data and machine learning at the ONS: developing innovative temporal disaggregation techniques

Geoffrey Megardon, Andrew Sutton, Ciaren Taylor, Daniel Ollerenshaw, Simona Astadurova, Tingting Peng, Philip Wales
Office for National Statistics

The ONS is undergoing an important transformation driven by the leitmotif: better data for better decision. An important milestone of the transformation is to understand how combined administrative data sources (as opposed to our survey data alone) and data science methods can improve our output statistics. With this goal in mind, the Economic Microdata Research team is creating a longitudinal database gathering different administrative data sources. One recurring problem that we face is producing monthly outputs from mixed, often lower frequency data. We are working on VAT returns data from HMRC, which consist of the turnover and expenditure of all the VAT registered businesses in the UK. We use a temporal disaggregation method on quarterly returners that is constrained by using a relevant group of monthly returners as indicators.Currently, the Standard Industry Classification (SIC) is assumed to be a good grouping to use when constructing indicators. Our work explores two ways to improve upon using SIC: 1) using hierarchical clustering so that the structure of the data itself is used to infer the best indicators, 2) using the high-dimensional Trade data from HMRC in combination to the VAT data to get a better measure of the similarity between two businesses. We will be presenting our methods, and the different challenges we faced while working with a high-dimensional, mixed and large dataset. We will show comparisons of our results against those obtained using SIC.

4.7 Contributed - Data Science: Misc

Wednesday 4 September 9am

On Simulating Ultra High-Dimensional Multivariate Discrete Data

Alfred Schissler
University of Nevada, Reno

In this era of Big Data, it is critical to realistically simulate data to conduct informative Monte Carlo studies. This is often problematic when data are inherently multivariate while at the same time are (ultra-) high dimensional. This situation appears frequently in observational data found on online and in high-throughput biomedical experiments (e.g., RNA-sequencing). Due to the difficulty in simulating realistic correlated data points, researchers often resort to simulation designs that posit independence --- greatly diminishing the insight into the empirical operating characteristics of any proposed methodology. A major challenge lies in the computational complexity involved in simulating these massive multivariate constructions. In this paper, we first review high-dimensional multivariate approaches and discuss relative merits of the approaches. Then we propose a fairly general procedure to simulate high-dimensional multivariate distributions with pre-specified marginal characteristics and a covariance matrix. Finally, we apply our method to simulate RNA-sequencing data sets (dimension > 20,000) with heterogeneous negative binomial marginals.

4.7 Contributed - Data Science: Misc

Wednesday 4 September 9am

Performance of calibrated non-positive definite covariance matrix under a classification problem

Ronald Wesonga
Department of Statistics, College of Science, Sultan Qaboos University

Calibration methods to deal with covariance matrix via matrix nearness problem are explored. The classification problems is studied to assess correct classification rates (CCR) in the presence of the data missing at random (MAR) mechanism. We explore some calibration methods. Given an arbitrary square matrix X of order n, the Frobenius norm is used. The nearest symmetric semi-definite matrix P0 (X) = Qdiag{max(λ1, 0), ..., max(λn, 0)}QT is assessed against another calibrated form Pc(X) = Qdiag{max(λ1,c),...,max(λn,c)}QT to evaluate information recovery from the former. A novel algorithm will be presented to deal with the problem of CCR in the presence of MAR problem using the calibrated covariance matrix. Results based on simulations and actual data will be presented and discussed.

4.9 Contributed - Environmental / Spatial Statistics: Climate

Wednesday 4 September 9am

A bivariate spatiotemporal model to estimate the occurrence of ozone and PM10 emergency alerts in Mexico City

Eliane R. Rodrigues[1], Philip A. White[2], Alan E. Gelfand[2], Guadalupe Tzintzun[3]
[1] Universidad Nacional Autonoma de Mexico (UNAM), [2] Duke University, [3] Instituto Nacional de Ecologia y Cambio Climatico

Mexico City is among the many cities suffering from high levels of pollution. In order to reduce population exposure and, therefore, the health hazard that is a consequence of this exposure, Mexico City has implemented a series of preventive measures. Among them are emergency alerts which are declared whenever high levels of ozone and/or PM10 occur. In the present talk a bivariate spatiotemporal model is considered to predict local pollution emergencies and to assess compliance to Mexican ambient air quality standards. Hourly ozone and PM10 measurements from 24 stations across Mexico City from 2017 are analyzed. Using this model, future pollutant levels using current weather conditions and recent pollutant concentrations may be predicted as well as the regional maxima needed to estimate the probability of future pollution emergencies. We discuss how predicted compliance to legislated pollution limits varied across regions within Mexico City in 2017. We find that predicted probability of pollution emergencies is limited to a few time periods. In contrast, we show that predicted exceedance of Mexican ambient air quality standards is a common, nearly daily occurrence.

This is a joint work with Philip A. White, Alan E. Gelfand, and Guadalupe Tzintzun.

4.9 Contributed - Environmental / Spatial Statistics: Climate

Wednesday 4 September 9am

A new statistical approach to forecasting non-stationary climate indices

Philip Sansom, Daniel Williamson, David Stephenson
University of Exeter

We present a new statistical model for forecasting non-stationary climate indices based only on observations and compare it to state-of-the-art dynamical seasonal forecast models and existing methods for estimating potential predictability. The new model separates a time series of a climate index in to long-term trends (including multi-decadal signals), seasonal behaviour, day-to-day variability, and observation errors. Simple non-stationary statistical models are proposed for each component so that the state of each can be simultaneously estimated at any time. The rate of change in each component is also estimated directly from the time series. Periods of enhanced predictability due to transient changes in either the mean or day-to-day autocorrelation of the climate index can also be identified and exploited for prediction.We apply the proposed statistical model to the problem of seasonal forecasting for the North Atlantic Oscillation (NAO). The statistical model identifies a period of enhanced predictability of the NAO between December and March, extending into November and April, due to transient changes in the mean of the NAO which we attribute to external forcing. Forecasts of the winter (Dec-Jan-Feb) mean NAO achieve a correlation with the observations greater than 0.5, making the statistical model comparable to the state-of-the-art GloSea5 seasonal forecast model. The forecast skill is actually greater in spring (Mar-Apr-May) than in winter. Limited forecast skill is also found in summer and autumn due to the presence of long-term trends.The statistical model can also be used to study potential predictability in climate indices. We estimate that around 60% of the inter-annual variability in the winter mean NAO is attributable to an externally forced signal, 30% to accumulated day-to-day variability, and 10% to long-term trends. These estimates are surprisingly close to those from earlier potential predictability studies which confounded the predictable signal with changes in the long-term mean.

4.9 Contributed - Environmental / Spatial Statistics: Climate

Wednesday 4 September 9am

Quantifying uncertainty in climate projections based on emergent constraints

Philip Sansom
University of Exeter

We present a new Bayesian statistical model and web-based tool for quantifying uncertainty in climate projections based on emergent constraints. Emergent constraints are a popular but controversial method of constraining projections of future climate based on simulations from ensembles of climate models and historical observations.We explain the statistical assumptions that lead to the standard model for emergent constraints. We highlight the different sources of uncertainty present when exploiting emergent constraints. We argue that the standard model for emergent constraints underestimates the uncertainty in projections by ignoring uncertainty about how constraints estimated from ensembles of climate models will manifest in the real world. We propose a new Bayesian statistical model for emergent constraints that accounts for this missing uncertainty. The new statistical model is a simple generalisation of the standard model for emergent constraints that allows for more realistic judgements about how emergent constraints will manifest in the real world. The new statistical model has been implemented as a web-based tool that enables quick and simple visualisation of the effects of accounting for the missing uncertainty. We demonstrate both the statistical model and the accompanying web-based tool by application to emergent constraints on equilibrium climate sensitivity.

Keynote 4 - Barnett Lecture

Wednesday 4 September 10.10am

Data lakes from the data deluge- a digital environment vision

Marian Scott
Glasgow University

The landscape of environmental data comprises outputs from process models, data generated from earth observation, and from in situ (static and mobile) sensor networks, and citizen science contributions including from social media feeds- so everything is anchored in time and space (where we may need to account for sampling effort, bias, detectability). In many instances, we measure proxies of the environmental variable of interest, and often, we still persist in thinking of the environmental media in silos. Learning from these many data sources is challenging, our objectives may be prediction so a black box might be sufficient, but in terms of explainability and transparency, will not not adequate. Variation, variation, and more variation clouds our perceptions of the state of the environment (driven by data quality, by measurement uncertainty and by natural variation, we know each particle of soil is different at a fundamental level, each parcel of air or water similarly different). Our understanding of fundamental environmental process may be high-level and profound, but may include detailed biology, chemistry or geology, so that expert knowledge is contributing a further form of data. Are we foolish to think we can learn the rules that govern our world from our imperfect observations? In this presentation, I will reflect through a number of examples on what we might mean by digital environment and what that vision is.

5.1 Medical: Precision Medicine

Wednesday 4 September 11.50am

Dynamic modelling of single-case (n-of-1) data: challenges and novel applications

Rute Vieira
University of Aberdeen

Introduction: Single-case studies are increasingly recognised as a valid and efficient mechanism for making individualized evidence-based treatment decisions. Statistical analyses of N-of-1 data require accurate modelling of the outcome variable while accounting for its distribution, time-related trend and error structures (e.g. autocorrelation) as well as reporting readily usable effect sizes for clinical decision making. A substantial number of statistical approaches have been documented but no consensus exist on which method is most appropriate for which kind of design and data.

Methods: In this paper we discuss, from a statistical perspective, N-of-1 studies and describe a regression method for the analysis of N-of-1 data, borrowing ideas from longitudinal and event history methodologies which explicitly incorporate the role of time and the dependence of future on past. The aims include identifying predictors of response, describing adaptive changes over time, or predicting future behaviour given prior history.

Results: Different applications of the method are presented, specifically to identify predictors of physical activity during retirement transition and predictors of photoprotection behaviour in patients with Xeroderma Pigmentosum, a rare inherited condition characterized by an extreme sensitivity to ultraviolet rays from sunlight. Our approach was shown to be flexible, adaptable to different types of outcomes and capable with dealing with the different challenges inherent to N-of-1 modelling.

Conclusions: N-of-1 studies are feasible and helpful in identifying individualised predictors of behaviour and facilitating the development of individualised intervention plans. Dynamic modelling has the potential to expand access of N-of-1 researchers to robust and user-friendly statistical methods.

5.1 Medical: Precision Medicine

Wednesday 4 September 11.50am

Predicting disease progression in neurodegenerative diseases with high phenotypic variability

Frank Dondelinger
Lancaster University

Identifying factors that influence the clinical progression of neurodegenerative diseases is of critical importance to both experimentalists trying to understand the disease mechanisms, and clinical researchers trying to develop improved therapies. While much effort has gone into the detection of risk factors for a given disease, most of these approaches ignore the inherent variability in the clinical phenotypes. We have developed a high-dimensional mixture model approach for jointly solving the problem of data-driven estimation of clinical phenotypes and prediction of disease progression. Longitudinal dynamics are captured via a mixed model approach, and we take into account both the distribution of the response and the distribution of the covariates for estimating the disease phenotypes. We demonstrate the performance of our method by applying it to data from the PROACT database on amytrophic lateral sclerosis, as well as data from the Alzheimer's Disease Neuroimaging Initiative (ADNI, Mueller et al., 2005). We show that in both cases joint inference of the subtypes and predictors improves the prediction performance, and hence the clinical usefulness of our results.

5.1 Medical: Precision Medicine

Wednesday 4 September 11.50am

Methodological challenges for precision public health

David Wright[1], David Taylor-Robinson[2], Frank Kee[1]
[1] *Queen's University Belfast,* [2] *Institute of Population Health Sciences, University of Liverpool*

Objectives: Considerable debate surrounds the usefulness of applying precision medicine techniques to the sphere of public health. We highlight some of the methodological challenges which must be addressed to ensure that the benefits of “precision public health” outweigh the harms.

Challenges: Building an evidence base for use of genomic or detailed digitally collected phenotypic data in public health faces three main challenges. First, to identify a target for intervention, risk of a given outcome must be defined for each individual. Widely used epidemiological methods focus on estimating population level parameters which do not translate readily at the individual level. Current analyses often fail to account for complex, interdependent influences on individual health or for reporting biases, residual confounding, collinearity and measurement error when assessing risk factors. Second, there is a need to adopt new methods to design and test the effectiveness of individual level interventions. Conventional parallel group randomised controlled trials are insufficient; n-of-1 studies and micro-randomised trials offer potential solutions. Intervention design should be theory-based and clearly define what represents a clinically important difference for each individual. Third, a framework is needed to assess unintended consequences and potential harms of precision public health evaluations. As more people use “wearable” technologies we are likely to face more “incidental” findings that have limited “actionability” leading to unnecessary interventions. Allowing apps and AI to substitute for insightful professional intervention in support of behaviour change may open the door to commercial interests and risks of harm. Most psychological studies of the impact of biological or behavioural risk profiling have focussed on single behaviours - risks of information overload or inducing adverse compensatory behaviours have generally not been studied.

Conclusions: There is a need for statisticians to engage with these challenges to allow robust evaluation of precision public health interventions before they are implemented more widely.

5.1 Medical: Precision Medicine

Wednesday 4 September 11.50am

Robust joint modelling: a new approach to handle time-varying outlier impacts

Laura Boyle[1], Özgür Asar[2], Lisa McFetridge[3], Jonas Wallin[4]
[1] *The University of Adelaide,* [2] *Acıbadem Mehmet Ali Aydınlar University,* [3] *Queen's University Belfast,* [4] *Lund University*

Joint modelling approaches, which simultaneously analyse a longitudinal and survival process, can be used to understand the relationship between repeated measurement data and patient survival in medical research. Medical data commonly contain outlying values, which can occur for a number of reasons, such as patients who are experiencing a period of stabilising in response to a new treatment. To handle such data, robust joint modelling approaches have been developed, utilising t-distributional assumptions to down-weigh the detrimental impact of longitudinal outliers on parameter estimation. However, these approaches do not currently account for scenarios where the pattern of outliers varies over time.This research advances joint modelling methodology by allowing the degrees of freedom parameter for the longitudinal residuals to vary through time using a natural cubic spline. Results from a simulation study will be presented, in addition to application of the method to two motivating medical datasets –a cohort of patients with liver disease, and a cohort of renal patients from Northern Ireland. Through comparison of a range of robust joint models, this research not only stresses the need to properly account for longitudinal outliers, a practice not commonly done in the literature, but highlights the bias from not properly estimating the degrees of freedom parameter. Each technique presented can be fitted using the R software package “robjm”.

5.2 Official & Public Policy: Feeding the beast - Satisfying user demand for precision, timelines and speed!

Wednesday 4 September 11.50am

How does a Government Statistical Agency respond to the challenge of developing and monitoring a societal wellbeing framework?

Niall O'Neill
NISRA

Part 1: A Wellbeing Framework for Northern Ireland - How does a Government Statistical Agency respond to the challenge of developing and monitoring a societal wellbeing framework?The Draft Programme for Government (2016-21) and Northern Ireland Civil Service (NICS) Outcomes Delivery Plan set the strategic context for central government in Northern Ireland, while Community Development Plans set the context in our local government districts. A wellbeing framework of 12 Outcomes and 49 population indicators overarches all of this work.The framework is heavily dependent on the application of official statistics produced by NISRA and, with data at its core, NISRA statisticians in all government departments are responding to the increased user demand and focus that the framework has brought.This presentation will cover NISRA's role in the development of the framework, including work to date in addressing gaps where data for proposed indicators did not exist. We will also consider how NISRA has met the challenges of reporting change over time through the establishment of a Technical Assessment Panel.

5.2 Official & Public Policy: Feeding the beast - Satisfying user demand for precision, timelines and speed!

Wednesday 4 September 11.50am

How NISRA has been informing EU Exit discussions

Chris Ganley
NISRA

The purpose of the presentation is to provide a brief overview of some of the work NISRA have been undertaking over the last 2 years to inform the BREXIT debate. A number of projects have been undertaken in order to answer queries from policy colleagues around issues such as:What is the value of trade with the EU?Who are our biggest trading partners?What is the profile of NI businesses involved in trade with the EU?What products do we sell to / buy from X, Y, Z?What impact would a change in the tariff regime have on NI trade?Who's most likely be impacted by changes to trading arrangements? How many people cross the NI-ROI border?How many vehicles cross the border?How many EU Nationals work in NI?How much freight is crossing the Irish Sea?Our work has fed into HM Government Position papers, internal briefings for policy colleagues in the Department for the Economy, Senior Civil Service and Department for Exiting the EU. We have also published slide packs and papers to make the data accessible to the public to ensure that there are informed debates around BREXIT.

5.2 Official & Public Policy: Feeding the beast - Satisfying user demand for precision, timelines and speed!

Wednesday 4 September 11.50am

Forecasting Emergency Care Occupancy

Eugene Mooney, Kieran Taggart, Michael O'Donnell
NISRA

It is well known that our Emergency Care Departments are experiencing significant challenges and pressures, as a result of increased demands and tighter budgets. Despite the wealth of information routinely published informing health care professionals and the wider public of these pressures, it has been of little use in assisting them to better plan their services.This presentation will outline how NISRA statisticians have utilised this existing information to forecast occupancy levels in emergency departments during each hour up to 6 weeks in advance, and enable Health & Social Care Trusts to better plan their unscheduled care services.It will also demonstrate the use of technological developments that allow this information to be almost completely automated and presented in a user friendly format, whereby the user can adjust the information for their own specific needs.

5.3 Application of Statistics: Data in the NHS - facilitating quality improvement in healthcare

Wednesday 4 September 11.50am

Monitoring perinatal mortality in the UK

Bradley Manktelow
University of Leicester

Audit data plays an important role in patient safety and quality improvement in healthcare. Care providers (e.g. hospitals, surgeons) are able to compare their clinical performance to other providers, specified benchmarks or their own previous performance. It is important that such information is presented in a manner that allows clear interpreted and can facilitate the improvement of the quality of care.Over the past 15 years, funnel plots have become a standard graphical method with which to present summary statistics for individual healthcare providers. Derived from SPC control charts, funnel plots are constructed to show the value of the summary statistic together with control limits indicating the 'expected' range of these values. Usually the control limits are selected to allow a precise probabilistic interpretation. However, the use of funnel plots raises several issues, including the plausibility of the assumptions and the use of statistical criteria to identify clinical outliers.One existing alternative approach is to use multilevel models to estimate empirical Bayes estimates of the values of the summary statistics. This approach can potentially allow the inclusion of clinical criteria into the identification of outliers.These two methods will be described and illustrated using mortality data from MBRRACE-UK, the national surveillance programme for stillbirths and neonatal deaths in the UK. The potential advantages and challenges of the methods will be discussed, together with the interpretation of the summary statistics.

5.3 Application of Statistics: Data in the NHS - facilitating quality improvement in healthcare

Wednesday 4 September 11.50am

Linked cluster-randomised trials of audit and feedback: A “split-block” design

Rebecca Walwyn
University of Leeds

Feedback interventions are frequently tailored across a variety of healthcare settings. The AFFINITIE programme evaluated feedback interventions used in audit and feedback cycles embedded within a UK national comparative audit programme of blood transfusion practice. Two feedback interventions (“content” and “follow-on support”) were evaluated in the context of both a surgical audit and a haematology audit in two linked cluster-randomised, 2x2 factorial trials, one for each audit topic. The national audit data was used as baseline and outcome data. As eligible, consenting NHS trusts across the entire country were involved in both trials, we considered whether trusts should maintain their allocations across trials or be re-randomised to interventions for the second trial. Patients and health professionals were largely separate across the trials. Researchers were interested in the size of the intervention effects in each setting, but also in whether these effects generalise across settings. Design of experiments literature has developed largely independently of clinical trials. Multi-stratum designs include cluster-randomised, split-plot and split-block designs as special cases. Intervention packages, or selected components, are randomised to clusters, possibly at more than one level, and outcomes are measured on patients. We will comment on the issues around using national audit data in clinical trials. We will then illustrate a range of possible multistratum designs, using our motivating example, and discuss the implications for equipoise, contamination, balance, convergence and power of using a split-block design to understand the generalisability of interventions across settings as part of a single research programme. Finally, we will outline the limitations of existing multistratum designs for use in a clinical trials context.

5.3 Application of Statistics: Data in the NHS - facilitating quality improvement in healthcare

Wednesday 4 September 11.50am

Quasi-Experimental designs for quality improvement research in Acute Ischemic Stroke Care

Abdel Douiri
King's College London

The decision-making process in healthcare is still predominantly driven by descriptive analytics which only superficially explore the information contained within the data. This talk will focus on methods for stroke quality of care improvement and will endeavour to provide strategies and tools to implement the use of statistical methods and designs (e.g: quasi-experimental design) to identify performance gaps, suggest possible solutions and inform and evaluate quality improvement initiatives. A study example using propensity scores and national real-world data to evaluate Intravenous thrombolysis with alteplase, one of the few approved treatments for acute ischemic stroke care, will be presented and challenges will be discussed.

5.5 Methods and Theory: Theoretical advances in experimental design

Wednesday 4 September 11.50am

Design and analysis of experiments testing for biodiversity

R. A. Bailey
University of St Andrews

It is now widely believed that biological diversity is good for the environment. One way that ecologists test this is to place random collections of species in mini-environments and then measure some outcome. Others use a carefully chosencollection of species.Is the outcome affected by the number of different species present, or is it just the number of members of each species that matters? And is there an interaction between these? Are the answers affected by the temperature, or by the complexityof the environment?I have been working with a group of fresh-water ecologists on the design and analysis of such experiments. Our subsets of species are carefully chosen, not random. We design the combinations of these subsets with levels of other factors.We also fit a nested family of plausible models to the data. Our results suggest that the underlying model is not diversity at all. One of my crucial inputs has been the use of Hasse diagrams as a way of understanding a complicated family of plausible models for the expectation of the response.

5.5 Methods and Theory: Theoretical advances in experimental design

Wednesday 4 September 11.50am

Construction of Blocked Factorial and Fractional Factorial Designs

Janet Godolphin
University of Surrey

Generating matrix methods are used to construct 2n and 2n-p factorial designs with single and double confounding, that is, with one or two forms of blocking. For designs with one form of blocking, the construction of single replicate designs in blocks of size 2q, enabling estimation of all main effects and maximising the number of estimable two factor interactions is demonstrated. For the situation in which a specific subset of the two factor interactions is of interest, methods on proper vertex colouring from Graph Theory are exploited to inform the generator matrix and yield a suitable design.Designs with two forms of blocking are represented as one or more rectangular arrays with rows and columns corresponding to blocking factors. Templates are given for generating matrices which yield single replicate constructions to estimate all main effects and the maximum number of two factor interactions. The construction approaches are extended to accommodate fractional designs.

5.6 Communicating & Teaching Statistics: Show me the stats

Wednesday 4 September 11.50am

Show me the question

Andrew Wright
Novartis

Graphics are often designed “bottom up”. That is, we start from the data and decide how best to display it. However, since a graph is most impactful if it is designed to answer a specific question, it is more meaningful to start from the question itself. To this end, a “top down” approach is suggested. We call this approach “Question-Based Visualizations” (QBV). The QBV approach to developing visualizations encourages teams to first align on the key questions they want to address with their data. Then, and only then, it is decided how to best create a graph (or set of graphs) to answer this question. This subtle yet important shift in the process of graph creation greatly improves the impact of the visualizations. In cases where a question is best answered by a combination of information from disparate dataset domains (e.g., vital signs and biomarkers), this is also more easily done “top down”; the “bottom up” approach tends toward separate displays for separate domains. Visualizations that are developed using this question-based approach will be illustrated in two different settings: a static graphic for inclusion in a report and interactive graphics embedded within a Shiny app.Putting the questions front and center when developing visualizations of our data can help us to better influence our teams.

5.6 Communicating & Teaching Statistics: Show me the stats

Wednesday 4 September 11.50am

Look at graphs

Allan Reese
Independent consultant

Why, despite many sources of good advice including Tufte's books and a tutorial by Vandemeulebroecke et al (to appear in CPT:PSP), do bad graphs continue to appear in print and during presentations? A corollary of the assumption "graphs communicate" is that people glance at them and move on. Secondly perhaps, "familiarity breeds contempt". When I devised a university module in Graphical Interpretation of Data (GID), colleagues in the Statistics department were happy to agree to assist at practical sessions: "I've used graphs all my career so I'm familiar with them." Once they attended, the general response was, "I've never thought about graphs that way before!" Students, from a variety of disciplines, often signed up in the expectation of learning some snappy tricks with Excel. They had been beguiled by advertising along the lines of, "Buy our software and with a few clicks you'll be creating presentation masterpieces." How do we improve? How long do you/should you spend looking at an individual graph? Should you start with the overall pattern or the detail – the wood or the trees? The central premise of GID is that there are no fixed rules, no absolutes. As a communication medium, better graphs – and better understanding of what you are doing with them – come from practice and reflection. The lack of revision (redrafting) is, I suggest, a major reason for poor graphs. Graphs exist in the context of their data and analyses. Many published graphs appear to have come straight from software with all the default settings. The surrounding text has been reviewed and redrafted many times – why not the graphs?In this short session, I will critique a few example graphs to emphasize the GID approach. I look forward to comments from the panel and audience.

5.6 Communicating & Teaching Statistics: Show me the stats

Wednesday 4 September 11.50am

Say one thing: numbers in the news

Robert Cuffe
BBC

When the aim is to communicate with millions at a glance, graphics need to be clear. Robert will share examples from BBC online and TV graphics and some insight from audience research about what works.

5.7 Data Science: Professional Ethics in Data Science

Wednesday 4 September 11.50am

How the ASA Ethical Guidelines for Statistical Practice comprise data science and "data ethics"

Rochelle Tractenberg
Georgetown University and the Collaborative for Research on Outcomes and -Metrics

The American Statistical Association (ASA) first established its Ethical Guidelines for Statistical Practice in 1999. In 2014, the ASA Committee on Professional Ethics set out to revise these Guidelines. These were revised in 2016 and 2018. Lessons learned included how to balance "completeness" with shortness of attention spans; the need for applicability of the Guidelines to any person who works directly or indirectly with data; and the fact that Guideline principles may conflict in any given case, so rather than "learn the Guidelines", ethical practitioners need to learn how to use the Guidelines. Recent worldwide interest in data science and data ethics creates challenges for teaching statistics and statistical ethics, but also creates new "data analysts" who are not, and who would not identify themselves as, "professional statisticians". If applicability of the Guidelines is subtly shifted from "ethical statistical practice" to "ethical quantitative practice", it can comprise data science and data ethics.

5.7 Data Science: Professional Ethics in Data Science

Wednesday 4 September 11.50am

Data Science: Professional Ethics in Data Science

Leone Wardman
ONS

Data science methods are increasingly common within analytical professions and come with unique ethical challenges. The Royal Statistical Society (RSS) Data Science Section and the Institute and Faculty of Actuaries (IFoA) collaborated to explore the practical and ethical implications of using data science in accordance with their respective strong professional values, and overarching commitment to the public interest. In 2018, joint regional workshops were held with a broad range of data science professionals to discuss key questions for data science, including the responsibility of data scientists to society and practices for a good data science 'workflow'. A joint focus group considered the findings from the workshops, along with a range of existing data ethics frameworks relevant to data science and Artificial Intelligence, to develop globally applicable ethical and practical guidance for our respective members when using data science. This session provides a background to the RSS and IFoA collaboration and an exciting insight into the proposed joint ethical guidance. In particular, it explores five overarching ethical themes and corresponding working practices to assist members in working ethically with data science.

5.8 Business, Industry & Finance: Embedding Data Analytics in the Business Process

Wednesday 4 September 11.50am

Using data to become a clearer, quicker and tougher regulator

Johanna Hutchinson
The Pensions Regulator

Increasing poverty in old age, a growing retired population and a fluctuating economy have led to considerable pension reforms in government. High profile cases of deficits in pension scheme from BHS, Carillion, Tata Steel and Toys R Us, alongside the launch of obligatory work-based pension schemes have kept the Pensions Regulator in the UK's media. The proactive regulation of pension schemes has required a new approach, an investment in technology and a dependency on data. In this presentation I will discuss how we built data-driven regulatory processes, provided operational efficiencies and new insights using data science, data management and data governance techniques.

5.8 Business, Industry & Finance: Embedding Data Analytics in the Business Process

Wednesday 4 September 11.50am

Uncertainty in real-world applications of AI

Dongho Kim
PROWLER.io Limited

Real world applications of AI often present challenges such as uncertainty, and multiple agents with different goals operating in dynamic environments. In order to mitigate this, we at PROWLER.io pursue a principled machine learning approach combining probabilistic modelling and decision theory.In this talk, we will describe some examples of applied AI where there's not only uncertainty, but also unknown values that need to be worked with. We motivate the need for an integrated approach that combines several sub-disciplines to solve problems such as scheduling transport when stock levels are unknown and unobservable.

5.8 Business, Industry & Finance: Embedding Data Analytics in the Business Process

Wednesday 4 September 11.50am

IceCAM – the Iceberg Crop Adaptive Model – Minimising food waste by adapting growing programmes to the weather

Iain Flint
G's Growers

Some fresh produce like lettuce and celery cannot be stored long-term and so must be sold the week it is harvested. This means growers must anticipate how temperature, light and other environmental factors will affect their crops to sow the right amount of crop months in advance to meet their sales programme on any given week. To solve this, G's Growers have developed a crop growth model – IceCAM – that reliably predicts harvest date in response to the weather in a given year for G's main crop Iceberg lettuce, as well as Gem and Romaine lettuce, Spinach, and Radish. This allows the growers to target higher yields with the confidence that they will meet their sales programme, meaning less land and resources are required to produce the same amount of food. This talk will discuss the progress of the model so far as well as recent developments including development of a biological crop model, learning from remote sensing images, and Monte Carlo simulation of forecast weather conditions to make optimal programme decisions under uncertainty, as well as future ambitions for the project.

5.9 Environmental / Spatial Statistics: Climate Change

Wednesday 4 September 11.50am

Bayesian Additive Regression Trees for palaeoclimate reconstruction

Andrew Parnell
Maynooth University

Learning about past climate is of key importance in discovering how fast climate can change. Most reconstructions of past climate use proxy data (pollen, coral, tree rings, etc) which are treated as highly noisy weather stations. Statistical methods are required to calibrate the proxy data to learn about their relationship to climate, and also to model the way that climate might change over time.A large number of different statistical methods have been proposed to perform reconstruction, across the full gamut of statistical techniques. In this talk I will conduct a brief review of some of these and then propose a new one based on the method of Bayesian Additive Regression Trees (BART). The key development includes that of extending the BART model to multivariate responses with non-ignorable missingness. No previous knowledge of BART is necessary, but attendees should be familiar with basic Bayesian model building and fitting for e.g. linear regression models.

5.9 Environmental / Spatial Statistics: Climate Change

Wednesday 4 September 11.50am

Statistics for Heatwaves and Extreme Waves in a Changing Climate

Jonathan Tawn
Lancaster University

Understanding how environmental risk factors change over time is vital for forward planning and infrastructure design. If interest is in the behaviour of heatwaves or extreme wave events in 50 years' time, standard extreme value methods using only observational data cannot be applied with any confidence given the level of change in climate that is anticipated over this time scale. Additional information from climate models is essential. However, using global climate models induces complications linked to their low spatial resolution, the choice of climate change scenario, and the sensitivity to climate model formulation. We will illustrate the problems of dealing with these features for estimating risk by presenting new downscaling methods for extreme wave events and a framework for heatwave modelling which allows for different climate change effects on the marginal distribution and duration properties of heatwaves. In our applications to North Sea waves and French heatwaves we find that there could be a significant increase in risk, with extreme events increasing at a faster rate than mean values.

6.1 Medical: Bridging public understanding of health and data sharing

Wednesday 4 September 2.20pm

Communicating about patient data - some experience from Understanding Patient Data

Natalie Banner
Understanding Patient Data

Understanding Patient Data is an independent initiative hosted at Wellcome that seeks to make the uses of patient data more visible, understandable and trustworthy. There are enormous potential benefits from using data collected in routine care better to improve health, but many people also understandably have concerns about their privacy, and also want to ensure data is used for public and social good. The language around data is often technical, complex and inconsistent, with unclear rules across organisations and a lack of transparency over what happens to patient data and why. UPD aims to help demystify the use of data, providing accessible tools and resources for clinicians, researchers, and anyone who wants to create trustworthy systems and talk about data use to patients and the public. In this talk, Natalie will give an overview of UPD's work to date, resources produced and collaborations with other organisations, and explore some of the challenges of engaging with people about data use in accessible ways.

6.1 Medical: Bridging public understanding of health and data sharing

Wednesday 4 September 2.20pm

Socialising expertise: people-centred data governance of health information

Stephanie Mulrine, Madeleine Murtagh, Mwenza Blell, Joel Minion, Mavis Machirori
Newcastle University

Across the UK and internationally, steps are being taken to revolutionise access to patient data collected by health services like the NHS. Work is underway to provide and assemble the necessary legal and technical expertise. Yet without public and patient perspectives on the use of their data, the potential for unsuccessful, dangerous or discriminatory consequences is a real risk.Focus group participants in the North East of England and North Cumbria were presented with information about a proposed technological solution: the Great North Care Record. A clear ethical framework of values emerged: respect, reciprocity, fairness, agency, privacy, transparency and trust. Whilst there is broad public support, these findings demonstrate concerns need to be addressed and a set of key values which must be adhered to. Further work (currently underway) to develop a co-production approach will inform the design and governance of the evolving process of health information exchange and access.

6.1 Medical: Bridging public understanding of health and data sharing

Wednesday 4 September 2.20pm

The patient imperative to 'use MY data'

Debbie Keatley[1], Finian Bannon[2], Hannah McKenna[2]
[1] *use MY data,* [2] *Centre for Public Health, Queen's University Belfast*

use MY data is a movement of patients, carers and relativesuse MY data supports and promotes the protection of individual choice, freedom and privacy in the sharing of healthcare data to improve patient treatments and outcomes. use MY data endeavours to highlight the many benefits that appropriate usage of healthcare data can make, to save lives and improve care for all.use MY data aims to educate and harness the patient voice to understand aspirations and concerns around the use of data in healthcare delivery, in service improvement and in research, aimed at improving patient decision making, treatment and experience.Our vision is of every patient willingly giving their data to help others, knowing that effective safeguards to maintain the confidentiality and anonymity of their data are applied consistently, transparently and rigorously.Debbie will talk about the growth of the movement and current topics in data access and usage, highlighting problems in access to routinely collected patient data, and how there may be too much risk-aversion to giving access to data.Debbie will then highlight what is possible when data is used. Using the example of routes/pathways to diagnosis she will share findings from the recently completed Northern Ireland Pathways to Cancer Diagnosis study and will demonstrate how Routes to Diagnosis studies in England have been a driver for positive change; demonstrating the benefits of collecting and using patient data.

6.5 Methods & Theory: Statistical sparsity

Wednesday 4 September 2.20pm

Overfitting correction in multivariate survival analysis

Anthony Coolen
King's College London

Generalized linear multivariate models are ubiquitous in applied statistics. They include logistic regression for binary classification, proportional hazards (Cox) models for time-to-event data or ordinal class data, frailty models, and so on. In most of these models, parameters are estimated using Maximum Likelihood (ML) regression. However, when the data dimension p is comparable to the sample size N, as is very often the case in post-genome medicine, maximum likelihood inference breaks down due to overfitting. This prompted the introduction of the so-called regularized models, which are equivalent to point-estimate approximations of Bayesian parameter inference (MAP regression), and require undesirable hyper-parameter tuning. Overfitting causes not only excess noise that is not captured by conventional p-values, but it also induces a strong inference bias, leading to consistent over-estimation of associations. In this presentation we show principles and applications of a recent formalism for modelling quantitatively the effect of overfitting in generalized linear models with ML or MAP regression. This formalism is based on the so-called replica method from statistical physics. It leads to precise formulae for the correction of multivariate regression parameters and base hazard rates in the overfitting regime, and for analytical formulae for optimal regularization hyper-parameters. In this presentation we will focus mainly on its applications in multivariate survival analysis.

6.6 Communicating & Teaching Statistics: Can mathematics anxiety obstruct learning statistics in university students?

Wednesday 4 September 2.20pm

An interactional and inclusive approach to enhancing students' engagement with statistics.

Meena Mehta Kotecha
The London School of Economics and Political Science

The world is increasingly driven by data science with statistics becoming increasingly important in all areas of society. Statistical skills are pivotal in the global employment market and university degree programmes are becoming progressively quantitative, involving statistical analysis for virtually all disciplines. It is imperative that we design and deliver statistics courses for non-specialists bearing this in mind, to prepare our students to face today's data driven work environment. I found mathematics anxiety (MA) (Richardson, & Suinn, 1972) rather than statistics anxiety (SA) (Cruise, Cash, and Bolton, 1985) to be an obstructing factor to non-specialists' engagement with statistics courses. Approximately 78% of non-specialists report MA (Kotecha, 2016). Aversive prior experiences with mathematics, poor achievement in mathematics and a low sense of mathematics self-efficacy have been found to be meaningful antecedent correlates of SA (Zeidner, 1991). MA is found to be an antecedent of SA and both forms of anxiety are found to be highly correlated (Paechter et al. 2017). I developed a research informed intervention to reduce MA reported by university non-specialist university students. A mixed methods approach was used in this research. The findings reflect a highly significant reduction in MA reported by non-specialist university students after the intervention was implemented. While this result is promising, it also raises important research questions that require rigorous inquiry. As Merton (1948) wrote "More is learned from a single success than from the multiple failures. A single success proves it can be done. Thereafter, it is necessary only to learn what made it work." My current research at the University of Cambridge is aimed at gaining an in-depth insight into the evident reduction in MA and developing a theory that integrates the use of social media -which is an important feature of the intervention used since September 2012 to present. This talk should trigger thought provoking questions and discussion.

6.6 Communicating & Teaching Statistics: Can mathematics anxiety obstruct learning statistics in university students?

Wednesday 4 September 2.20pm

Are graduate students scared of statistics? Statistics Anxiety: barriers, enablers and policies.

Carlos Fresneda-Portillo
Oxford Brookes University

Statistics anxiety, defined as `the feelings of anxiety encountered when taking a statistics course or doing statistical analyses; that is gathering, processing and interpreting' (Cruise, Cash, & Bolton, 1985), has become one of the global sustainable challenges (UN Sustainable Development Goals, 4.c). Given that the growing demand on data analysis experts and the fact that career choices are influenced by mathematics and statistics anxiety, it becomes of paramount importance to gain deeper understanding about what factors influence Statistics Anxiety. Research points out that as many as 80% of graduate students may be suffering from Statistics Anxiety (Onwuegbuzie, 2004). In this talk, we will aim to provide an overview on where we are with Statistics and Mathematics anxiety, and what educational policies are underway for amelioration.

6.6 Communicating & Teaching Statistics: Can mathematics anxiety obstruct learning statistics in university students?

Wednesday 4 September 2.20pm

How do medical students feel about statistics and data skills during a 10-week individual project and what support do they access?

Jamie Sergeant
University of Manchester

At the University of Manchester, undergraduate medical students undertake a 10-week student-selected project in their third year. Projects can either be “research” or “non-research” (e.g. service evaluation or clinical audit) and the aims for all projects include the critical review of literature, understanding research methodology and understanding basic statistical analysis. Support available to students to help them meet these aims includes lectures, statistical advice appointments and signposting to resources. But how confident and well-prepared do students feel to tackle the statistical aspects of a project and does this influence their choice of project? Once they have completed their project, do students feel that their statistics and data skills have been enhanced? What about their level of confidence? What sources of statistical support were students aware of and which did they access during the project? This talk will describe an attempt to answer these questions by asking students.

7.1 Contributed - Medicine: Clinical Trials

Wednesday 4 September 4.10pm

Optimal treatment allocations in sequential multiple assignment randomized trial (SMART) design

Mirjam Moerbeek, Andrea Morciano
Utrecht University, the Netherlands

In an adaptive treatment strategy (ATS) subjects are assigned to a sequence of treatments. The treatments that are available in a given stage depend on the previous treatments and whether the subject responded to these treatments. A sequential multiple assignment randomized trial (SMART) combines multiple ATSs with the aim to select the best one. Consider as an example a SMART for overweight patients: they are either randomized to a dietary intervention or a physical activity intervention in the first stage. Those who lost a sufficient amount of weight in the first stage keep their intervention in the second. Those who did not are randomized to either the other intervention or a combination of both interventions.Equal randomization is often used in SMARTS, but this is not necessarily the most efficient choice. The aim of the current study is to find optimal treatment allocations under either a fixed total sample size or a fixed budget. It uses multiple objective optimal design methodology to take into account multiple pairwise comparisons among ATSs, and to put a weight of importance on each comparison. Optimal treatment allocations depend on the response rates to current treatments, and these response rates are often not known a priori. For that reason we derive robust optimal design using maximin methodology. We compare our optimal design with equal randomization and show how much efficiency is lost by using the latter. Our optimal design methodology is implemented in a Shiny app.

7.1 Contributed - Medicine: Clinical Trials

Wednesday 4 September 4.10pm

Do mixed neuropathologies affect cognitive decline and possible Alzheimer's Disease Clinical Trial failure?

Sumali Bajaj[1], David X. Thomas[2], Kevin McRae-McKee[1], Christoforos Hadjichrysanthou[1], Frank De Wolf[1], John Collinge[2], Roy M. Anderson[1]
[1] *Imperial College London,* [2] *University College London*

Background: Phase 3 clinical trials of potential disease modifying therapies for Alzheimer's Disease (AD) have all been unsuccessful to date. AD is characterized by ß-amyloid (Aß) plaques and neurofibrillary (Tau) tangles in the brain. While it is widely known that AD is a multifactorial disease and that patients frequently have other pathologies in the brain, the extent to which these co-occurring neuropathologies (TDP-43, CAA and Lewy Body) can affect cognitive decline is poorly understood at present.

Methods: We developed a statistical model for describing the effect of the three co-occurring pathologies in patients with AD and non-AD pathology on the progression of cognitive scores (Mini-Mental State Examination, MMSE; Clinical Dementia Rating Scale Sum of Boxes, CDR-SB) over time. Longitudinal scores were modelled within a Bayesian framework using a hierarchical regression model with random intercepts and slopes. We also performed beta regression as part of our sensitivity analysis. Beta distribution was appropriate for our bounded continuous, skewed outcome and heteroskedastic data. Posterior predictive checks were performed for model validation.

Results: In AD individuals we found that TDP-43 and CAA co-pathologies were significantly associated with a faster rate of cognitive decline, after adjusting for confounders. Having TDP-43 pathology, even without AD, can drive cognitive decline at a substantially fast annual rate (ß = -1.19; 95% CI: -1.71, -0.69 and 0.70; 95% CI: 0.35, 1.06 for MMSE and CDR-SB respectively). We demonstrate a high prevalence of moderate-severe co-pathologies (63% in our data had TDP-43 or CAA along with AD pathology) and show their effect on cognition. These results suggest that even if a drug completely removed all parenchymal Aβ and Tau neuropathology within a well-defined AD population, a high proportion of individuals may continue to cognitively decline, potentially being another possible explanation for the failure of clinical trials solely target ß Amyloid.

7.1 Contributed - Medicine: Clinical Trials

Wednesday 4 September 4.10pm

Using Statistical Modeling to Identify the Useful Surrogate Outcomes in Critical Care Studies

Rejina Verghis[1], Bronagh Blackwood[1], Cliona McDowell[2], Daniel McAuley[1], Mike Clarke[12]
[1] Queens University Belfast, [2] Northern Ireland Clinical Trials Unit

Introduction: In critical care, many confirmatory studies fail, and this is sometimes because the outcome measures used in the earlier studies were not valid. This contributes to research waste and the aim of this research was to identify and validate surrogate outcome measures in critical care trials.

Methods: A systematic review generated a full list of outcome measures used in critical care trials, which was shortened based on the frequency of the reports and clinical judgement. Critical care clinicians and researchers were then surveyed to identify the importance of the identified outcomes to their clinical decision making and research. The association between the surrogate outcomes and mortality was tested using Cox and joint modelling on relevant secondary data.

Results: Forty eight articles were purposively sampled until data saturation, to generate a list of 247 different outcomes. This was reduced to a smaller list of 22 outcomes, which were broadly classified as resource use outcomes, disease severity scores, routinely collected physiology outcomes and biomarkers. SOFA, a disease severity score in critical care, was one of the most frequently reported outcomes and was considered an influential outcome by clinicians and researchers. In the HARP2 study, the hazard ratio (95% CI), for SOFA score, was1.54 (1.18 to 2.01) and 1.38 (1.29 to 1.48) for the cox model and joint model respectively.

Conclusions: A variety of outcome measures have been used in critical care research, and these are analysed and reported in many different ways. The variability makes comparisons between studies in systematic reviews time consuming and difficult or impossible but the SOFA score seems promising. In the dataset tested, the joint model estimates were more precise than the cox model estimates and these analyses will be repeated on additional datasets, to investigate the change in SOFA and associated difference in mortality rate.

7.2 Contributed - Official & Public Policy: Children's health and wellbeing

Wednesday 4 September 4.10pm

Understanding the histories of children in care and links to outcomes with longitudinal analysis.

Cecilia Macintyre[1], Gillian Raab[2]
[1] *Scottish Government,* [2] *University of Edinburgh*

‘Looked after children' [1]are children in the care of their local authority. There are many reasons children may become looked after: they face abuse or neglect at home; they have disabilities that require special care; they are unaccompanied minors seeking asylum, or who have been illegally trafficked into the UK; or they have been involved in the youth justice system. Some children are cared for at home, with regular contact with social services. In other cases the child or young person is cared for in placements away from their normal place of residence by foster or kinship carers, prospective adopters, or in residential care homes, schools or secure units. The experience of children and young people in care vary in terms of the length of time and the range of different placements. This presentation considers the link between and a child’s experience in care and educational outcomes derived from administrative data. The data on looked after children is based on annual returns which are made to Scottish Government by local authorities[2]. Individual level data has been collected since 2008 with details of each placement. The data on absence, attendance and exclusions [3]is collected every second year from publicly funded schools.Data returns covering a ten year period were combined, resulting in over 50,000 children and around 120,00 placements. The analysis of school attendance of children, included explanatory factors of length and consistency of placement, age and gender. Challenges which were addressed included incomplete linkage due to missing identifiers , and selective effects of data due to incomplete coverage of care history for children born prior to 2008 when data collection started.The presentation will discuss how the Scottish Government has worked with partners – local authority, academic and other public bodies - to improve the quality of the data used in this project

[1] Children (Scotland) Act 1995

[2] https://www2.gov.scot/Topics/Statistics/Browse/Children/PubChildrenSocialWork

[3] https://www2.gov.scot/Topics/Statistics/ScotXed/SchoolEducation/AttendanceAbsenceExclusions

7.2 Contributed - Official & Public Policy: Children's health and wellbeing

Wednesday 4 September 4.10pm

Predictors of Mental Disorders in Children. Analysis of the Mental Health of Children and Young People in England, 2017

Jodie Davis, Tim Vizard
Office for National Statistics

This session presents the methods and findings from a logistic regression analysis of the 2017 Mental Health of Children and Young People (MHCYP) survey in England. The relationships between characteristics of a child (factors) and the development of mental disorders are complex. More than one factor may be associated with a child having a disorder, and these factors may also be associated with each other. Our research aimed to identify which factors were associated with mental disorders in children (2 to 16 years old), after controlling for other factors at the same time. By utilising a backwards logistic regression, the research explored associations between fourteen factors and the presence of mental disorders in children. Factors were grouped into demographic, family and socioeconomic characteristics of children, with separate models run for preschool, primary school and secondary school aged children.The results from this research were published by NHS Digital in March 2019, in a report authored by the Office for National Statistics (ONS). The findings showed that factors such as parental mental health, family functioning and receipt of welfare benefits were associated with mental disorders in children of all ages and disorder types. However, the research also found some factors were associated with particular age groups and disorder types. This session, presented by ONS, will provide:an overview of methods adopted in this analysis, some of the key findings from the research, and future analysis plans for the survey.The 2017 Mental Health of Children and Young People Survey was commissioned by NHS Digital, and conducted by ONS, the National Centre for Social Research and YouthInMind.

For more information on the survey, please visit: https://digital.nhs.uk/data-and-information/publications/statistical/mental-health-of-children-and-young-people-in-england/2017/2017.

7.2 Contributed - Official & Public Policy: Children's health and wellbeing

Wednesday 4 September 4.10pm

Why our our children lonely? Findings from multivariate statistics and in-depth interviews with 10-15 year olds

Ellie Osborn, Charlotte Hassell, Eleanor Rees
Office for National Statistics

Loneliness is normal and transient, but physical, mental and social problems arise when it becomes chronic. It has a profoundly negative effect on health at all ages and Hawkley and Cacioppo (2010) called it “social equivalent of physical pain”. Loneliness is associated with poor health outcomes for younger ages as much as for adults. As part of the Prime Minister’s strategy to overcome loneliness, the Social Analysis branch in ONS undertook research exploring factors associated with loneliness among children. We developed interim National Measures of Loneliness and noted that there is currently much more extensive and robust data available on loneliness in older people, but comparatively little for children (aged 10-15) and young people (aged 16-24). We have also focused on analysing new data collected to address this gap to map out what loneliness in these ages looks like in the UK. Following the published exploratory analysis last year, regression analysis has now been run to further understand the relationship between the factors highlighted. Focusing on children, we have been able to further assess the relationship between loneliness, poverty, the impact of a child’s perception, and their parent’s life satisfaction. We have also revisited the interviews to use the children’s own words to complement this.

We will focus on:

- Which factors are associated with children’s experiences of loneliness
- What children say in their own words about their experiences of loneliness
- How we can ensure children’s loneliness is addressed to avoid illness in later life
- How we work with charities and policy-makers to ensure our analysis improves loneliness in children as well as older ages

The analysis is key for the strategy for loneliness, and ONS works closely with DCMS and charities to ensure that the research has the greatest impact.

7.3 Contributed - Medical: Data Science I

Wednesday 4 September 4.10pm

Applying Statistical Learning to Nuclear Magnetic Resonance Metabolic Profiling to Predict Spontaneous Preterm Birth

Juhi Gupta[1], Angharad Care[1], Bertram Müller-Myhsok[12], Laura Goodfellow[1], Ana Alfirevic[1], Zarko Alfirevic[1], Lu-Yun Lian[1], Marie Phelan[1]
[1] *University of Liverpool,* [2] *Max Planck Institute of Psychiatry (Germany)*

Preterm birth (PTB) is defined by the World Health Organisation as birth prior to 37 weeks of gestation. Complications of spontaneous preterm birth (sPTB) are the leading cause of death in children under age 5. Associated risk factors include genetics, lifestyle, or infection. Understanding the mechanism of sPTB can aid development of novel methods to prevent PTB. This study aims to investigate the metabolome of PTB high-risk patients to identify potential biomarkers, using statistical methods. Serum samples were collected at 16 and 20 weeks of gestation from “high risk” women (with a history of sPTB or preterm premature rupture of membranes (PPROM) under 34 weeks gestation), and “low risk” parous women (with previous births delivered at term). Pregnancy outcomes were categorised into sPTB, PPROM, high-risk term (HTERM) and low-risk term (LTERM). Serum samples (n = 647) were acquired using solution-state 1D 1H Nuclear Magnetic Resonance (NMR) technology (Bruker Avance III). Vendor software (Topspin 3.1) was used for processing and quality control of spectra with Chenomx and in-house standards for annotating and identifying over 40 different metabolites. MetaboAnalyst web tool and R programming were used to analyse the spectral bins.Probabilistic neural network analysis using a single timepoint between cases and controls, resulted in a predictive AUC of 0.89 (LOOCV). One-way ANOVA calculated across the 4 phenotypic sub-groups, resulted in a subset of metabolites discriminating between HTERM:LTERM, sPTB:LTERM and PPROM:LTERM ($p < 0.05$ post-hoc=Tukey HSD). Metabolites included phenylalanine and acetate, both reported in current PTB biomarker literature.Variability due to exogenous substances and environmental factors, such as drug intake or diet, require further investigation. Future work includes integration of metabolomics with genomics, transcriptomics and proteomics. Investigating multiple omics data for this patient cohort will allow for a better understanding of the pathophysiology involved in initiation of early labour.

7.3 Contributed - Medical: Data Science I

Wednesday 4 September 4.10pm

Interpreting Clinical Narrative Diagnosis Models with Sentence Importance

Mark Ormerod, Jesús Martínez del Rincón, Neil Robertson, Bernadette McGuinness, Barry Devereux
Queen's University Belfast

Despite advances in the application of deep neural networks to various kinds of medical data, extracting information from unstructured textual sources remains a challenging task. Using a dataset of de-identified clinical letters gathered at a memory clinic, we evaluate recurrent neural networks (RNNs) on their ability to predict patients' diagnoses of 'Dementia', 'Mild Cognitive Impairment' or 'Non-impaired'. This classification framework can also have applications in the automatic identification of patients as candidates for clinical trials.After showing that models trained on state-of-the-art sentence-level embeddings outperform both word-level models and a recent benchmark model that fine-tunes a pre-trained general-domain language model, we probe sentence embedding models in order to reveal interpretable insights into the types of sentence-level representations the RNNs build. Specifically, we take a measure of sentence importance with respect to a given class and identify clusters of sentences in the embedding space that correlate strongly with importance scores for each class. Extracting the most frequent phrases within each group of sentence representations shows that the model is sensitive to sentences that cluster around semantic concepts such as a patient's level of geriatric depression and how independent the patient is in their daily activities.In addition to showing which sentences in a document are most informative about the patient's condition, our method can identify the types of sentences that can lead the model to make incorrect diagnoses.

7.3 Contributed - Medical: Data Science I

Wednesday 4 September 4.10pm

A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models

Evangelia Christodoulou[1], Jie Ma[2], Gary S. Collins[23], Ewout W. Steyerberg[4], Jan Y. Verbakel[567], Ben Van Calster[45]
[1] KU Leuven, Belgium, [2] Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Botnar Research Centre, University of Oxford, [3] Oxford University Hospitals NHS Foundation Trust, [4] Department of Biomedical Data Sciences, Leiden University Medical Centre, Leiden, the Netherlands, [5] KU Leuven, Department of Development & Regeneration, Leuven, Belgium; [6] KU Leuven, Department of Public Health & Primary Care, Leuven, Belgium; [7] Nuffield Department of Primary Care Health Sciences

Objective: To compare performance of logistic regression (LR) with machine learning (ML) for clinical prediction modeling, and to describe methodology and reporting.

Methods: We conducted a Medline literature search (1/2016 to 8/2017), and extracted comparisons between LR and ML models for binary outcomes. We classified comparisons as at low or high risk of bias, where the latter was defined as: incorrect or unclear validation, or different methodology for LR vs ML in terms of number of predictors, variable selection, or validation.

Results: We included 71 out of 927 studies. The median sample size was 1250 (range 72-3,994,872), with 19 predictors considered (range 5-563) and 8 events per predictor (range 0.3-6,697). The most common ML methods were classification trees (30 studies), random forests (28), artificial neural networks (26), and support vector machines (24). Sixty-four (90%) studies used the area under the receiver operating characteristic curve (AUC) to assess discrimination. Calibration was not addressed in 56 (79%) studies. In 48 studies (68%), we observed unclear reporting or potential biases in validation procedures for one or more algorithms. We identified 282 comparisons between a LR and ML model (AUC range, 0.52-0.99). For 145 comparisons at low risk of bias, the difference in logit(AUC) between LR and ML was 0.00 (95% confidence interval, -0.18 to 0.18). For 137 comparisons at high risk of bias, logit(AUC) was 0.34 (0.20 to 0.47) higher for ML.

Conclusions: Methodology and reporting in studies that compare LR and ML needs improvement. We found no evidence of superior performance of ML in clinical prediction modeling. Future research should focus on delineating the type of predictive problems in which different algorithms have maximal value.

7.4 RSS Prize Winners: Statistical Excellence Award for Early Career Writing 2019

Wednesday 4 September 4.10pm

The flying bomb and the actuary

Luke Shaw[1], Liam Shaw[2]
[1] *Office for National Statistics (ONS),* [2] *Nuffield Department of Medicine, University of Oxford*

In the early morning of 13th June 1944, one week after the D-Day landings, the Nazi regime launched a new weapon at London. The first Vergeltungswaffe 1 (Vengeance Weapon 1, or V-1), hit a railway bridge in Mile End, killing six people and leaving 200 people homeless. Over the following 9 months, more than 2,300 ‘flying bombs’ fell on London, killing an estimated 5,500 people. The V-1 was the first ever operational cruise missile, capable of delivering an 850kg warhead from a range of 250km at any time, day or night. It was certainly a formidable weapon of terror, but if it could be accurately targeted at munitions factories it would also become a strategic threat to the war effort. As the campaign progressed, it was clear that the V-1s sometimes fell in clusters: was this the result of random chance, or of precision guidance? The answer to this question was of critical importance. Humans are good at seeing patterns, but statistical analysis can help us decide whether those patterns are due to chance. Were the V-1s really targeted in clusters, or were they falling at random? Answering this question means going on a journey through bomb maps, statistics textbooks, and even a novel, following in the footsteps of a British actuary.The authors recreated R.D. Clarke's classic textbook example of the Poisson Distribution, using V-1 impacts, 75 years after Clarke did and found similar results. This was through one author (Liam) meticulously adding all the impacts to a Google Maps layer from bomb damage maps republished in 2015. They also performed some extra assessment of randomness using bootstrapping, and found a less significant p-value when using all data available; with North West London less-heavily bombed. Finally, they assessed V-2 impacts, and looked into the history of Thomas Pynchon's use of the Poisson Distribution within his novel "Gravity's Rainbow".

7.4 RSS Prize Winners: Statistical Excellence Award for Early Career Writing 2019

Wednesday 4 September 4.10pm

Polarisation and the popular vote in Switzerland

Maximilian Aigner
Université de Lausanne

The recent use of referendums in European representative democracies (Hungary 2016, UK 2016, Catalonia 2017) has highlighted fears regarding populism, but also led to greater interest in direct democracy. In general, popular votes are often uncontroversial, but those that generate significant attention and polarise voters often show 1) high participation and 2) a close outcome (small margin). To highlight such votes, we combine both characteristics into a polarisation score which is related to the Gini inequality index.Using data from Switzerland, a country with direct-democratic features, we apply this score to the last hundred years' worth of referendums, highlighting the most polarising votes of each decade. Summarising values by their maximum leads to a well-developed mathematical theory in many situations, and we use results from Extreme Value Theory to study the probability and time to recurrence of a Brexit-sized event in Switzerland.The results indicate that events equally or more polarising than the 2016 UK referendum are quite frequent. However, 'populist' uses of referendums remain a common feature even in a country with a long tradition of direct democracy.

7.4 RSS Prize Winners: Statistical Excellence Award for Early Career Writing 2019

Wednesday 4 September 4.10pm

A story about a tiny bot

Marco Antonio Andrade Barrera
National Autonomous University of Mexico

This work's aim is to show a simple example of a working artificial intelligence application, a bot which can emulate a very specialized human thinking. It is created using Object Oriented Programming and a popular machine learning approach. All design steps are described, starting with basic elements like the bot's name and birth, following the data gathering and the machine learning algorithms, up to reach with a bot able to classify stock chart patterns with high accuracy and speed. The work ends pointing that currently, AI applications can provide us tools or abilities to do very complex processes or at least avoid time- consuming processes that are out of the human capabilities. In this respect, AI is improving human possibilities.

7.5 Contributed - Methods & Theory: Survival Analysis I

Wednesday 4 September 4.10pm

Goodness-of-fit tests for the cure rate in a mixture cure model

Ursula U. Mueller[1], Ingrid Van Keilegom [2]
[1] *Texas A&M University,* [2] *KU Leuven*

We consider models for time-to-event data that allow that an event, e.g., a relapse of a disease, never occurs for a certain percentage p of the population, called the cure rate. We suppose that these data are subject to random right censoring and we model the data using a mixture cure model, in which the survival function of the uncured subjects is left unspecified. The aim is to test whether the cure rate p, as a function of the covariates, satisfies a certain parametric model. To do so, we propose a test statistic that is inspired by a goodness-of-fit test for a regression function by Härdle and Mammen. We show that the statistic is asymptotically normally distributed under the null hypothesis that the model is correctly specified and under local alternatives. A bootstrap procedure is proposed to implement the test. The good performance of the approach is confirmed with simulations. For illustration we apply the test to data on the times between first and second birth.

7.5 Contributed - Methods & Theory: Survival Analysis I

Wednesday 4 September 4.10pm

On generalizing Banks' smoothed bootstrap method for right-censored data

Tahani Coolen-Maturi, Asamh Al Luhayb, Frank P.A. Coolen
Durham University

Bradley Efron invented the bootstrap method in 1979, which has become a popular tool ever since. In 1981, Efron introduced the bootstrap method for right-censored data, a particular type of data often encountered in reliability and survival analysis, where for a specific unit or individual it is only known that the event has not yet taken place at a specific time. In 1988, Banks introduced a histospline smoothed version of Efron bootstrap for real-valued data. The proposed method combines Banks' bootstrap and the right-censoring A(n) assumption introduced by Coolen and Yan in 2004. The right-censoring A(n) assumption provides a partially specified predictive probability distribution for a future observation based on n past observations. Banks' bootstrap is restricted to finite support; in this paper, we overcome this by assuming a distribution tail for the end interval(s). A simulation study has been conducted to compare the proposed method performance with Efron's bootstrap for right-censored data, where coverage probabilities of bootstrap confidence intervals for the three quartiles are compared. Our study shows that the proposed method performs better compared to Efron's method, in particular for small sample sizes.

7.6 Contributed - Communicating Statistics:

Wednesday 4 September 4.10pm

Civic Statistics: Big Ideas, Needs and Challenges. Why we need a new subdiscipline

Joachim Engel
Ludwigsburg University of Education

Effective citizen engagement with social issues requires active participation and a broad range of skills, including the understanding of data and statistics about society and our natural and social environment. We provide an overview of a subfield we call Civic Statisticsthat had been explored by a recent strategic partnership of six European universities under the Erasmus+ program of the EU. Civic Statistics is statistics about important societal trends and about topics that matter to the social and economic well-being of citizens. It sits at the crossroads of multiple disciplines including social science, education and statistics. Hence, a multidisciplinary educational perspective is needed, stepping outside the comfort zone of traditional statistics instruction.Understanding Civic Statistics is needed for participation in democratic societies, but involves data that often are open, official, multivariate in nature, and/or dynamic, which are usually not at the core of regular statistics instruction. Many statistics classes and educational curricula are not designed to teach relevant skills and improve learners' statistical literacy, despite the importance of engaging learners and future citizens with data about social issues and their connections to social policy. This talk gives an overview to characteristics of Civic Statistics and provides guidance to specific teaching materials developed by the ProCivicStat project.

7.6 Contributed - Communicating Statistics:

Wednesday 4 September 4.10pm

Insightful Analytics – surreptitiously adding thoughtful Statistics to the glamour of Data Science

Neil Spencer
University of Hertfordshire

Statistics is boring and old fashioned while Data Science is clever and glamourous. Perhaps not the majority view at an RSS conference but it is what much of the outside world sees. Degree programmes with titles associated with Data Science have mushroomed. In other degree programmes, modules with exciting Data Science-related titles are included to attract students.However, Statistics has not become irrelevant. While Data Science can provide amazing graphics based on enormous amounts of data, it still relies on Statistics to provide valid interpretation. While Data Science can build models that give excellent predictions, it still relies on Statistics to give insight into underlying mechanisms. While Data Science can classify cases using cunning algorithms, it still relies on Statistics to decide whether these have meaning.But if Statistics is so important, how do we get people to learn and use it without realising they are doing so? The same way we get dogs to take tablets – we trick them, perhaps by crushing them up and mixing with something we know they are going to gobble up. [N.B. the comparison between students and dogs is an affectionate one – dogs are fun-loving, friendly and capable of being trained to do amazing things.]In this talk we examine how a Statistics module can be transformed into a Data Science module in such a way that none of the important Statistical elements are side-lined. We surreptitiously address each Data Science topic in such a way that in order to overcome its shortcomings, Statistics needs to be understood and employed. We will also challenge audience members to suggest additional Statistical topics that can be hidden within a Data Science cloak. In time, we believe that the term “Statistics” will be recognised as that which gives real meaning to Data Science. However, in the meantime we avoid the “S” word and call it “Insightful Analytics”.

7.6 Contributed - Communicating Statistics:

Wednesday 4 September 4.10pm

Principles of statistical visualisation for public policy audiences: learning lessons from the past

Thomas King, Murray Dick[2]
-, [2] Newcastle University

Data visualisations that present data literally may be useful within a workflow or to communicate simple patterns, but they often have more aesthetic than illustrative value. More complex analysis requires visualisation of statistics derived from data, and many specialised practical tools exist for different substantive disciplines. However, statistical complexity abounds within niche policy areas, and so standard tools often do not exist; or where they do, the policy audience may be unfamiliar. We offer some principles that are based on theories of visual communication, for statistical visualisations aimed at public policy audiences, who know established regularities very well. These principles comprise three factors: The context in which information is communicated (pragmatics) that reduce cognitive load and contribute to the framing of the message.The content of these messages (semantics), concerning the statistical inferences designers wish to convey.The mode of convincing the audience (rhetoric); in this case, the application of visual metaphor towards embedding policy information (or disrupting prior knowledge or predispositions) amongst the audience.Features of effective communication are contrasted where others have failed despite being statistically sound; typically because their approach to visual communication lacked a sense of familiarity (pragmatics), policy-relevance (semantics), or a sense of sympathy (rhetoric) with their intended audience. We exhibit our own alternative examples based on historic data. Before formulating the message (the data visualisation) it is first necessary to reason through the communicative needs of the audience; in this case, key stakeholders in policy. This involves a process that contrasts with the broadcast approach of much data visualisation design. Once a robust inference is established, our approach is to identify a figurative representation of the data, in order to capture context, content and a mode of convincing, in order to persuade even audiences with prior knowledge, by means of visualisations that are sympathetic to their motivations.

7.7 Contributed - Data Ethics

Wednesday 4 September 4.10pm

The future of privacy and confidentiality methods

James Tucker, Keith Spicer
Office for National Statistics

A dramatic increase in the volume and sources of data presents an unprecedented opportunity to innovate. On the other hand, against a backdrop of increasingly sophisticated attacks by intruders and serious consequences of data breaches, the data revolution presents new challenges to protecting privacy and confidentiality, and ensuring that the associated methods are fit for purpose. It is vital that the statistical community understands and addresses these evolving challenges to provide a solid foundation for innovation and better decision making to take place. This is not a straightforward task, and there is a need to draw upon the full range of expertise in this fast-developing field.The Office for National Statistics joined forces with leading experts in privacy and data confidentiality from across the world to explore the latest advances in methods, such as synthetic data and differential privacy. Emerging themes are discussed, including the implications of increased use of data linking, and the potential of intruder testing techniques. Also outlined are the challenges faced by statisticians in keeping pace with the latest technological developments, and areas of future research, such as the practical applications of machine learning and artificial intelligence.

7.7 Contributed - Data Ethics

Wednesday 4 September 4.10pm

The Minority Report: Fairness and Explainability in Machine Learning

Stuart Millar
Centre for Secure Information Technologies, Queen's University Belfast

This talk aims to raise awareness of bias in predictive algorithms and real-world impact. As opposed to a hardcore technical discussion, we'll consider machine learning from an ethical point of view and shift the conversation from our models doing things right to doing the right thing. Aimed at anyone working in machine learning, deep learning or data science in any capacity, plus those interested in wider ethics, it features short interesting case studies from Microsoft, Google, the Met and US police forces. Handling bias and explainability in our models will be the difference between further breakthroughs in using machine learning, particularly deep learning, in the real-world, or not. With media coverage in recent times intimating that predictive algorithms can lead to, for example, racial, gender or sexual discrimination, there presently is a real risk models will make decisions we can't explain, and thereafter the damage will be already done. We'll discuss how we can start to take responsibility for mitigating bias when engineering our models and algorithms, including what tangible frameworks might look like to do so, and as researchers where we can publish our work helping to solve problems in this area. Ultimately, if we are dealing with predictive algorithms that make decisions about human beings, we must be responsible, be kind and be good citizens all at the same time. In theory, our models should have the capacity to act without the bias and discrimination that has plagued society for generations, but it is down to us to ensure this it the case and prevent this prejudice from becoming embedded in our tech, deliberately or otherwise.

7.7 Contributed - Data Ethics

Wednesday 4 September 4.10pm

Data ethics in a changing and challenging global context

Peter Elias
Warwick University

In 2016 the OECD published a report on 'Research Ethics and New Forms of Data for Social and Economic Research'. The report (https://doi.org/10.1787/5jln7vnpxs32-en) compiled over a two year period by statisticians, researchers, policy makers and experts in research ethics from 12 countries and the European Commission, put forward 13 recommendations designed to encourage research and statistical cooperation across international boundaries. The report was endorsed by the OECD Global Science Forum, a major international body for promoting scientific collaboration. This presentation provides information about these recommendations and, crucially, examines what has happened as a result.

7.8 Contributed - Business, Industry & Finance: Forecasting currency rates and stock market indices

Wednesday 4 September 4.10pm

Forecasting Risk Measures Using Intraday Data in a Generalized Autoregressive Score (GAS) Framework

Xiaohan Xue
ICMA Centre, Henley Business School, University of Reading

A new framework for the joint estimation and forecasting of dynamic Value-at-Risk (VaR) and Expected Shortfall (ES) is proposed by incorporating intraday information into a generalized autoregressive score (GAS) model. The GAS model was proposed by Patton, Ziegel, and Chen (2017) to estimate risk measures in a quantile regression framework. The intraday measures considered in this study include four realized measures: the realized variance at 5-min and 10-min sampling frequencies, and the overnight return incorporated into these two realized variances. In a forecasting study, the set of newly proposed semiparametric models is applied to 4 international stock market indices: the S&P 500, the Dow Jones Industrial Average, the NIKKEI 225 and the FTSE 100, and is compared with a range of parametric, nonparametric and semiparametric models including historical simulations, GARCH and the original GAS models. VaR and ES forecasts are backtested individually, and the joint loss function is used for comparisons. The GAS-FZ-Realized models, especially the GARCH-FZ and GAS-1F extended with the realized variance at the 5-minute frequency, outperform the benchmark methods consistently across all indices and various probability levels.

7.8 Contributed - Business, Industry & Finance: Forecasting currency rates and stock market indices

Wednesday 4 September 4.10pm

Dynamic functional time series forecasting of foreign exchange implied volatility surfaces

Han Lin Shang
Australian National University

Functional time series have become an integral part of both functional data and time series analysis. Recently, important contributions to methodology, theory and application for the prediction of future trajectories and the estimation of functional time series parameters have been made. This paper continues this line of research by considering both static and dynamic functional principal component analyses to decompose a time series of functions, and by adapting univariate, multivariate and multilevel functional time series methods to forecast implied volatility surfaces in foreign exchange markets. On one hand, the multivariate functional time series method first standardizes all series, then combines them into a long vector for each time period. Via a functional principal component analysis, the extracted functional principal components and their associated scores can be obtained from a joint covariance structure consisting of all series. On the other hand, the multilevel functional time series method uses multilevel functional principal component analysis of aggregate and maturity specific data to extract the common trend and maturity-specific residual trend among different maturities. Both multivariate and multilevel functional time series methods are able to capture correlation among maturities, and they are shown to be more accurate than the univariate functional time series for producing short- to medium-term forecasts. To demonstrate the effectiveness of the proposed methods, we examine the statistical significance of the three methods in forecasting daily EUR-USD, EUR-GBP and EUR-JPY implied volatility surfaces for a number of maturity periods.

7.8 Contributed - Business, Industry & Finance: Forecasting currency rates and stock market indices

Wednesday 4 September 4.10pm

Forecasting Cryptocurrencies Volatility with NonGaussian Garch Models

Massimiliano Giacalone[1], Raffaele Mattera[1], Roy Cerqueti[2]
[1] *University of Naples "Federico II" - Department of Economics and Statistics,* [2] *University of Macerata*

Financial time series usually are characterized by leptokurtic distribution, heavy-tails, skewness, volatility clustering and heteroscedasticity. Therefore, in this case thenormality assumption for the innovation's distribution is violated and the development ofmethods to handle the stylized facts became an important issue to model financial timeseries. In this context, an important problem is to define useful and efficient statisticalmethods for estimating and forecasting volatility.Despite the literature about volatility models for cryptocurrencies is not well developed yet,cryptocurrencies themselves are one of the most evident and interesting examples of theparticular characteristic of financial time series. In particular, the previous literature tried tofind the best GARCH model ignoring the assumptions about the innovation distribution,even if the characteristic of non-normality and skewness of the cryptocurrencies returndistribution are well known among researchers.We provide a comprehensive study about volatility estimation and forecasting, in terms of market capitalization, studying the dynamics of exchange rate with US Dollar.In order to evaluate the specification of each model we considered the Akaike InformationCriterion (AIC), that is one of the most common criteria used in the previous literature whilefrom the point of view of forecasting accuracy we used the Mean Square Error (MSE) andMean Absolute Error (MAE).We conclude that, while from the point of view of goodness-of-fitwith the cryptocurrency data, the best fitting distribution are for all the considered exchangerates the t-student and the Generalized Error Distribution while from the point of view ofvolatility forecasts we can get the most accurate predictions implementing a SkewedGeneralized Error Distribution in estimating GARCH model.

Main References1. Katsiampa, P., 2017, Volatility estimation for Bitcoin: A comparison of GARCH models,Economics Letters, 158, 3-6.2. Mattera, R. & Giacalone, M. (2018). Alternative distribution based GARCH models forBitcoin volatility estimation, The Empirical Economics Letters, 17(11), 1283-1288.

7.9 Contributed - Environmental / Spatial Statistics: Health Factors

Wednesday 4 September 4.10pm

Area interaction point processes for bivariate point patterns in a Bayesian context

Glenna Nightingale[1], Janine Illian[2], Ruth King[1]
[1] *University of Edinburgh,* [2] *University of St Andrews*

In this paper we consider bivariate point patterns which may contain both attractive and inhibitive interactions. The two subpatterns may depend on each other with both intra- and interspecific interactions possible. We use area interaction point processes for quantifying both attractive and inhibitive interactions in contrast to pairwise interaction point processes, typically model regular point patterns. The ability to permit both attraction and repulsion is a valuable feature and allows for the modelling of different forms of interactions in a given community. The differentiation between intra- and interspecific interactions in one model accounts for the fact that the presence of a second species may ``mask'' or ``magnify'' existing intraspecific interactions. A Bayesian approach has been applied for estimating interaction parameters and for discriminating between eight competing research hypotheses. For the particular application to modelling the interactions of species in a highly biodiverse forest, this study reveals posterior support for an interspecific interaction of attraction between the two species considered and may serve to inform forest rehabilitation schemes relating to this forest. Overall, knowledge of the interactions of key species in any given forest would be invaluable to reforestation efforts if this forest is later ravaged by wildfires.

7.9 Contributed - Environmental / Spatial Statistics: Health Factors

Wednesday 4 September 4.10pm

Simulating disease control strategies for Bovine Tuberculosis (bTB) in Northern Ireland

Emma Brown, Adele Marshall, Hannah Mitchell, Andrew Byrne
QUB

Bovine Tuberculosis (bTB) is endemic to Northern Ireland's (NI) cattle population with herd incidence rates steadily rising from 5.12% in 2010 to 9.61% in 2017 (DAERA, 2018). The corresponding cost, comprising of testing and compensation for culled cattle, was approximately £44 million in 2017 alone (NIAO, 2018).There is a need for mathematical modelling of bTB in NI to inform how it will behave when control strategies are applied. This research aims to introduce potential control strategies, apply them to an epidemic model (constructed as part of this project and is the first for bTB in NI), and assess the change in bTB incidence levels.The epidemic model to be used assumes hosts in the disease system start as susceptible (S), become latently infected through exposure (E), and finally become infectious (I). This SEI model assumes homogeneous mixing between hosts with density dependent transmission to account for increasing herd densities. The SEI model was extended to allow bTB to spread through cattle movements and for interactions between cattle and wildlife reservoirs. Control strategies will be investigated where the model's values for the transmission, demography, and bTB-detection parameters will be decreased to represent biosecurity, reducing herd density, and improving the infection detection. Other perturbations include restricting cattle movements and wildlife reservoir interactions. Strategies will also be combined to assess if multiple measures will be more effective than singular tactics.Simulating bTB control strategies reduces the need for resource-intensive field trials and provides a clear comparison between multiple control measure combinations. Knowledge from this analysis will inform and optimise resource allocation for a future testing regime. In doing so, herd incidence rates for bTB will become more manageable with the long-term goal of achieving the European Union's Officially-Tuberculosis Free (OTF) status when incidence levels stay below 0.1% for 3 consecutive years.

7.9 Contributed - Environmental / Spatial Statistics: Health Factors

Wednesday 4 September 4.10pm

Investigating the importance of environmental factors to understand renal disease attributed to uncertain aetiology

Jennifer McKinley[1], Ute Mueller[2], Peter Atkinson[3], Ulrich Ofterdinger[1], Siobhan Cox[1], Damian Fogarty[4], Chloe Jackson[1]
[1] Queen's University Belfast, [2] Edith Cowan University, [3] Lancaster University, UK, [4] Belfast Health & Social Care Trust, Belfast, Northern Ireland

There are several factors which are known to cause renal disease including age, ethnicity and pre-existing medical conditions. However, the UK Renal Registry (UKRR) and other national systems record a noteworthy number of incidences of renal replacement therapy (RRT) attributed to uncertain or unknown aetiology. Some have been attributed to environmental factors. The World Health Organisation divides naturally occurring elements into three groups on their nutritional significance in humans: 1) essential elements; 2) elements which are probably essential; and 3) potentially toxic elements (PTEs). For Northern Ireland, the Standardised Incidence Rates (SIRs) for patients starting RRT between 2006-2016, as provided by the UKRR, show that the SIRs for uncertain aetiology are up to 8% higher than would be expected based on Northern Ireland's age-specific incidence rates. Our study uses regional geochemical soil and stream water data to investigate the relationship between the environment and end-stage renal disease. Due to the constraints of the closed (compositional) nature of geochemical data, log-ratio and log-contrast approaches are explored. Poisson regression analysis is used to investigate any potential relationship between observed cases of end-stage renal disease with unknown aetiology and environmental covariates (soils, stream waters and stream sediments). A Tweedie model is investigated to account for the zero-inflated nature of the dependent variables which include Super Output areas with no incidences of uncertain aetiology. The results indicate that elemental associations are essential to appreciating the potential role of environmental factors in the development of renal disease.

8.1 Contributed - Medical: Treatments and Interventions

Thursday 5 September 9am

Optimum weighting schemes when performing Matching-Adjusted Indirect Comparisons

Dan Jackson, Kirsty Rhodes, Mario Ouwens
AstraZeneca

Indirect comparisons and network meta-analyses are standard methods to compare treatments from multiple trials. A common situation is where a company has individual level data on its own trial but only has published aggregate level data on a competitor's trial. In this situation, Matching Adjusted Indirect Comparison (MAIC) is an established way to adjust for between-trial population imbalances (different distributions of observed patient covariates). This methodology uses a method of moments, that does not require individual patient data in the competitor trial, to estimate the propensity score using logistic regression. Patients in the trial for which individual patient data are available are then weighted so that a population adjusted indirect comparison can be performed. The choice of logistic regression is somewhat arbitrary in this modelling, and it is natural to consider alternatives. We developed a novel approach where we used the method of moments in conjunction with the propensity score regression model that results in the largest possible effective sample size. Analytical results are obtained for the simplest possible case where there is one matching covariate, but more generally numerical methods must be used to perform the necessary optimisation. We investigated the properties of our proposed method by applying it to a well-known (but artificial) example from NICE Technical Support Document 18, and we also performed a simulation study based on this example. Worthwhile gains in effective sample size can be obtained in situations where considerable adjustment is required. We conclude that alternative weighting approaches are worthy of consideration when performing a MAIC, and our "optimal" weighting scheme should be considered for this purpose.

8.1 Contributed - Medical: Treatments and Interventions

Thursday 5 September 9am

Efficient Computation for Evaluation and Comparison of Phase I Oncology Study Designs

Jun Takeda
Astellas Pharma Inc.

One of the main purposes for a phase I oncology trial is to determine the maximum tolerated dose (MTD), which is defined as the maximum dose with acceptable dose limiting toxicity (DLT). Typically, such a study includes a sequence of cohorts with pre-specified rules to select the next dose (or to stop the study) and determine MTD. There are a variety of oncology phase I study designs. Famous among them are the three plus three (3+3) design, Bayesian continuous reassessment method (CRM) and Bayesian optimal interval (BOIN) design. Because the selection of study design and the tuning of parameters are crucial to planning a study, there is a need for methodologies to evaluate and to compare among different designs and different settings of parameters.A common practice to evaluate a study plan can be summarized as follows: Firstly, several scenarios of the dose-response are assumed. Then data are simulated based on the study plan and whether a subject has a DLT or not is simulated with a random sample from the binomial distribution corresponding to the assumed dose-response scenario. Finally, simulations are summarized, e.g., with operational characteristics. One of the biggest issue in this procedure is its computational burden, especially when Markov chain Monte Carlo (MCMC) or numerical integration is performed in a Bayesian study design. Researchers also encounter this issue when comparing their newly proposed designs with already-existing ones.To mitigate such burden and to gain computational precision, we propose an improved procedure. In the proposed procedure, the binomial random sampling part is replaced with an exact probability calculation, possible paths of stochastic process are identified, and the number of performing MCMC is reduced with the likelihood principle. With the improvement of computational time, we also propose random generation of dose-response relationship in order to make an exhaustive comparison among study designs.

8.1 Contributed - Medical: Treatments and Interventions

Thursday 5 September 9am

Graphical representation and comparison of attributable fractions across multiple disease risk factors

John Ferguson
NUI Galway

Population Attributable fractions (PAF) estimate the proportion of disease cases that might be avoided if a disease-risk factor could be removed at a population level; for instance, the reduction in the prevalence of lung cancer if nobody smoked. They are useful epidemiological tools to compare disease burden across differing risk factors and help suggest possible targets for health-interventions. In this talk, I describe appropriate definitions and estimation approaches for PAF using the Neyman-Rubin causal model. Several novel visualizations are suggested, displaying various forms of PAF across multiple different risk factors (both discrete and continuous). It is hoped that these visualizations will help prevent misinterpretation of PAF, and may assist practitioners in their search of good risk-factor targets for health interventions.

8.2 Contributed - Official & Public Policy:Capturing inequality

Thursday 5 September 9am

Statistics to Ensure Welfare for Ethnic Minorities

Kevin Johansen

In 2008, the ministry with responsible for Sami affairs decided to establish an Expert Group to edit and publish statistics on Sami issues.Before that, there were hardly any statistics available to developing policies on Sami affairs to advance Sami culture, language and welfare programs initiated by the ministry or the Sami parliament.The aim of our project is to contribute to in-depth knowledge on Sami society through numbers and commentary. This will help policy and law makers gain a broader knowledge for making qualified decisions on policy issues affecting especially the Sami people.Sami people live in Norway, Sweden, Finland and Russia and are designated by the UN as indigenous people. Sami people altogether number count 60 000, although the numbers aren´t definitive and vary. Because of this, statisticians haven´t had a big interest or knowledge in developing statistics on Sami issues.Through this project, the Expert Group has written and edited a vast number of articles on Sami affairs focusing on several important topics such as education for Samis, Sami health services, gender perspectives, Sami languages, Sami culture, reindeer herding, fisheries and so on.Our user surveys indicate that the publications are a crucial tool for policy makers, bureacrats, researchers and the media in getting access to Sami data sources.Indigenous people have the same need for Data and Statistics as everyone else.

8.2 Contributed - Official & Public Policy:Capturing inequality

Thursday 5 September 9am

Improving Public Confidence in Gender Pay Gap Reporting - An Overview of the RSS's Recommendations

Nigel Marriott
Marriott Statistical Consulting Ltd

Gender pay gap data is now part of the public discourse on the employment and remuneration of women in the UK workforce. The data has the potential to highlight good and bad practice among employers and could be transformational in reducing the gender pay gap. However, the first two years of published data has uncovered a number of issues in the way the data is collected, calculated and uploaded by the 10,000+ organisations who are required to submit such data. To avoid such issues undermining public confidence in this data, the RSS has published 10 recommendations to improve the situation. These recommendations cover 4 themes:-1. Making the published figures more consistent and easier to understand2. Increase the accuracy of the data and reduce chances of erroneous data3. Protecting the integrity of the data and reduce opportunities for gaming4. Keeping the reporting burden on employers as light as possibleThis talk will explain why these 10 recommendations have been and the outcomes that the RSS hope to see if these are implemented. The RSS has drawn heavily on the work done by Nigel Marriott who first posted his thoughts in this article https://marriott-stats.com/nigels-blog/gender-pay-gap-data-and-12-ways-to-improve-it/ . It is hoped that statisticians will support these recommendations and publicise them widely to HR professionals, journalists and policy makers.PS - As of April 4th, the RSS recommendations have not been published. The expected publication date is late April.

8.2 Contributed - Official & Public Policy:Capturing inequality

Thursday 5 September 9am

Using tax data to better capture top earners in household income inequality statistics

Martin Shine[1], Peter Matejic[2], Dominic Webber[1], Richard Tonkin[1]
[1] *Office for National Statistics,* [2] *Department for Work and Pensions*

The Office for National Statistics (ONS) annually releases “Effects of Taxes and Benefits” (ETB) which examines household disposable income across the UK based on the Living Costs and Food Survey. An issue with survey data is that top earners are both simultaneously under represented, and they under-report their income (Burkhauser et al. 2018). Inaccurate recording of top earners will inevitably result in a misleading picture of inequality, and household income across the UK. As a consequence, ONS, in collaboration with the Department for Work and Pensions (DWP), have explored methods to address under-reporting and under-coverage in ETB. In developing the adjustment, it was decided to build upon the world-leading “SPI adjustment” developed by DWP for the “Households Below Average Income” (HBAI) series. The SPI adjustment supplements survey data with administrative tax information contained within the tax data set “Survey of Personal Incomes” (SPI) from Her Majesty’s Revenue and Customs (HMRC). The current SPI adjustment used by DWP adjusts the top earning pensioners and non-pensioners separately. The average incomes of individuals above a threshold in SPI is calculated, and this income is assigned to the relevant rich individuals, replacing their survey-data based income.ONS adjustments explore the effects of adapting the SPI methodology to vary the threshold above which incomes in ETB are replaced and explores the use of multiple income bands above the threshold, assigning varying incomes from the SPI data to the “rich” individuals rather than one income for all of them. ONS recently released "Using tax data to better capture top earners in household income inequality statistics", exploring the effects of applying this new method of adjustment. This will be followed by a joint paper later in the year between ONS and DWP, going into more detail on the effect of the adjustment. The adjustment method will be developed to be used in headline ETB statistics from 2018/19.

8.3 Contributed - Applications of Statistics: Applications on animals and plants

Thursday 5 September 9am

A Bayesian analysis of animal movement data

Colin Gillespie
Newcastle University

Many of us have been struck by the inherent beauty of animals moving collectively. Starlings gathering at dusk in huge numbers to perform the most mesmerising of ballets, the entire flock moving as if some fluid object. Fish forming tight milling structures in defence against predation. Though we understand the evolutionary benefits offered to individuals by group behaviour, little is understood about how these structures are formed.Much work has been invested in developing theoretical agent-based models which seek to explain emergent behaviour by interactions at an individual level. However, these models have largely only been verified with a comparison to empirical observation at a qualitative level, and a thorough quantitative comparison between field data and theory has been lacking.This study is based on observations of sheep herds. Using a DJI Phantom 3 drone equipped with a video camera, we filmed sheep movement. The locations and movement of the sheep were then extracted using computer vision methods. This data was then analysed using a mechanistic flocking model - a type of Vicsek model. The model was fitted using a hierarchical structure with a Bayesian framework.

8.3 Contributed - Applications of Statistics: Applications on animals and plants

Thursday 5 September 9am

An application of covariate-based constrained randomisation in livestock research

Andrew Mead[1], Amy Thomas[1], Jess Evans[1], Bruce Griffith[1], Mark Eisler[2]
[1] *Rothamsted Research,* [2] *University of Bristol*

The North Wyke Farm Platform (a BBSRC National Capability, BBS/E/C/000J0100) is a unique research facility for the study of grassland livestock systems. It consists of 15 hydrologically-isolated catchments, grouped into three 21ha farmlets which are managed under different grassland farming systems. The platform is highly instrumented, providing high-quality data on characteristics of water, air, soil, plants and animals, much at a high temporal resolution.Herds of beef cattle and sheep are a key resource, used both to assess the relative performance of the different farming systems and explore aspects of animal management. As much research considers individual animals as the experimental unit, with replication farming system treatments only replicated between years, a balanced allocation of available animals to different treatments is essential.Covariate-based constrained randomisation is a common design approach in cluster or group randomised clinical trials, where cluster- or group-based covariates are used to allocate units to treatments. The standard approach considers all possible permutations of units to (usually 2) treatments, assessing differences in the covariate values of units allocated to each treatment, and selecting at random from those permutations meeting a specified level of balance between treatments.We describe the application of covariate-based constrained randomisation to the allocation of 90 beef cattle across the three farmlets, accounting for variation in 3 continuous covariates (age, weight, growth rate) and balancing the allocation for 5 breed/gender combinations. Further complications included the late identification of 7 of the animals, allowing for differences between sires, and interest in splitting each herd of 30 animals into 2 comparable sub-herds. We assess how the standard approach can be successfully adapted to cope with more than 2 treatments, very large numbers of permutations, and these additional complications.

8.3 Contributed - Applications of Statistics: Applications on animals and plants

Thursday 5 September 9am

Predicting Temporal Mutant-Treated Arabidopsis Thaliana Gene Expressions

Susana Conde, Daphne Ezer
The University of Warwick

Plants respond to their environment in a time-of-day dependent manner, and genetic perturbations may influence how plants react to their environments. We are interested in developing a model that predicts how plants will respond to a combination of genetic and environmental perturbations. Specifically, our aim is to predict mutant-treated gene expression over time, given temporal profiles of gene expression of the same genes in other experimental conditions (mutant-control, wild-type-treated, wild-type-control).Our data set comes from a ribonucleic acid (RNA)-seq experiment done in Arabidopsis thaliana plants in which researchers collected gene expression data at 8 time points during a two-hour period surrounding the dawn. The plants were exposed to two different temperatures. Additionally, researchers measured gene expression in mutants of some genes known to be related to light sensing processes.We compare the effectiveness of a linear regression and a function-on-function regression models. The latter, which allows for the dependence structure of the longitudinal measures for either response and/or explanatory variables, is estimated with the boosting algorithm in the R FDboost package (Brockhaus et al 2015). The response variable is the functional gene expression in the mutant-treated condition, and the explanatory variables are the functional gene expressions of wild-control, wild-treated and mutant-control.We fit our models to a set of m = 672 genes found to be early-morning expressed DNA-binding proteins, after imputing missing values. Our results show that the functional regression model outperforms the linear regression one with respect to the observed cross-validated root mean squared error. However, neither approach considers phase variabilities (i.e. time warping), which will be the phocus of future research.

References: Brockhaus, S., Scheipl, F., Hothorn, T., and Greven, S. (2015). The functional linear array model. Statistical Modelling, pages 279–300.

8.4 Contributed - Medical: Data Science II

Thursday 5 September 9am

UNCERTAINTY IN DIAGNOSTIC TESTING WITHOUT GOLD STANDARDS

Nicholas Gray, Marco De Angelis, Scott Ferson, Ryan Jackson, Uchenna Oparaji, , , , ,
University of Liverpool, Institute for Risk and Uncertainty

Problems with classification appear in various fields, from structural health in engineering, supervised learning in computer science, and patient diagnosis in medicine. Diagnostic tests are often imperfect, yielding false positives and false negatives. Bayesian methods are often used to estimate the probability that a patient has a condition, given the results of a medical test, the statistical characteristics of the test and the prevalence of a disease. However, these methods cannot be used without assuming that there is a reliable way to determine the true classification. When designing diagnostic tests or classification algorithms, analysts often refer to the 'gold standard' of evidence to calculate the statistical characteristics of the new test. However, these gold standards can themselves be imperfect and have some errors associated with them. For instance, for patients suspected of having Giant Cell Arteritis there are several tests that can be conducted in order to detect the disease. However, there is uncertainty in all these methods, including the biopsy which is called the gold standard by some. We present a series of logical rules that should hold when considering methods that aim to account for the error in diagnoses based upon an imperfect gold standard. These rules logically define limits where the uncertainty is at its maximum and minimum level. Ensuring that the uncertainty is never less than the uncertainty of the gold standard, and the uncertainty of the result is irreducible when a gold standard is missing. We will discuss a method to account for the uncertainty using imprecise probabilities. This method makes use of the gold standard test sensitivity and specificity to update these statistics for the new test. We use Bayesian methods to calculate the uncertainty associated with the probability that someone has a disease when given a positive test result.

8.4 Contributed - Medical: Data Science II

Thursday 5 September 9am

Spatial statistical modelling of retinal images from patients with diabetic macular oedema

Wenyue Zhu[1], Jae Yee Ku[1], Yalin Zheng[1], Paul Knox[1], Simon Harding[1], Ruwanthi Kolamunnage-Dona[1], Gabriela Czanner[2]
[1] *University of Liverpool,* [2] *Liverpool John Moores University*

Images are an essential aspect in monitoring and predicting diseases, especially in clinical management of retinal diseases. Spatial context in retinal images is highly relevant but often under-utilised. Current approaches involving analysing each location separately or analysing all locations together, which ignore the possible spatial correlations.To investigate the appropriate statistical approaches to account for the suitability of spatial error specification. We proposed a statistical inference framework for retinal images, which is based on a linear mixed effect model with a spatial error structure. The spatial topography was explained via fixed effect and spatial error structures, and correlation between eyes from the same patient was explained through nested random effects. We compared our method with multivariate analysis of variance (MANOVA) in analysis of spatial retinal thickness data from a prospective observational study, the Early Detection of Diabetic Macular Oedema (EDDMO) study involving 149 diabetic participants at their baseline visit. MANOVA analysis suggested that the overall retinal thickness of eyes with maculopathy are not significantly different from eyes with no maculopathy (p=0.11), while our spatial framework can detect the difference between the two groups (p=0.02).Simulation was conducted to evaluate our spatial statistical model framework, which illustrate how spatial correlations can affect the inferences about fixed effects. The simulation results confirmed the increased power of the statistical inference, which demonstrate the advantage of spatial modelling to provide more powerful statistical inference with power increase from 88.1% to 95.3%, 88.9% to 100% for moderate or high noise correlations.Our spatial approach addresses the need of correct adjustment for spatial correlations in ophthalmic image, and it can be extended into prognostic or predictive modelling in other diseases or imaging technologies.

8.4 Contributed - Medical: Data Science II

Thursday 5 September 9am

Predicting severe complications after cardiac surgery: methods and challenges

Linda Lapp[1], Matt-Mouley Bouamrane[1], Kimberley Kavanagh[1], Stefan Schraag[2]
[1] *University of Strathclyde,* [2] *Golden Jubilee National Hospital*

With the aging population patients undergoing cardiac surgery are more complicated to manage than ever before. Up to 50% of heart surgery patients experience postoperative complications, which can have a detrimental impact on patients' quality of life and healthcare resources.In a previous study we compared various machine learning methods (random forest, AdaBoost, Gradient Boosting Model and stacking) at predicting severe postoperative complications based on variables recorded before surgery. Our results showed that AdaBoost has the best overall performance (AUC=0.731). To improve the previously developed prediction model, we are looking into including cardiac intensive care data into our model.The data for this study comes from Golden Jubilee National Hospital, Scotland, and includes 6834 patient records. Two clinical audit datasets are used: one including data recorded at the preoperative clinic, such as patient characteristics, comorbidities and information about surgery; another including patients' vital signs and laboratory results recorded at cardiac intensive care unit. The time-series data about vital signs are recorded every other minute, and laboratory results are received at least twice a day.Although the number of patients with severe complications is considerably low in this dataset (5.50%), the size of intensive care unit data is large: the median number of days stayed in the ICU for these patients was 5.5 days, as opposed to 22 hours for patients with non-severe complications. We will discuss various methods that are appropriate for our classification problem and the challenges that arise from analysing such vast amount of real-world health data.This work is part of a larger study developing a clinical decision support tool predicting severe postoperative complications in cardiac patients. Such support system could help a clinician to identify patients who are at risk of having complications in order to allocate resources or avoid high-risk treatments.

8.5 Contributed - Methods & Theory: Survival Analysis II

Thursday 5 September 9am

Fast Bayesian hazard regression under general censoring via monotone p-splines

Matthias Kaeding
RWI - Leibniz Institute for Economic Research

The baseline hazard is the major building block of the Cox model, giving the instantaneous rate of failure, conditional on survival up to time t and covariate values of zero. Most Bayesian nonparametric approaches model the log-baseline hazard, causing the need for numerical integration for likelihood evaluation. We propose to model the integrated baseline hazard of the Cox model via monotone penalized B-splines instead; giving an analytically available likelihood, speeding up inference and eliminating approximation error. Left, right and interval censoring can be accounted for. The advantages of Bayesian P-splines carry over to the monotone case: (1) Fully automatic smoothness parameter estimation, (2) sparseness of involved design matrices. The closed form expression for the cumulative baseline hazard is used to extend inference beyond marginal effects on the hazard rate; allowing fast computation of the conditional mean and survival function and the simulation of random deviates. Inference is carried out using MCMC and posterior mode estimation. The proposed approach is tested using a Monte Carlo simulation and applied on a large data set of times until change of gas price.

8.5 Contributed - Methods & Theory: Survival Analysis II

Thursday 5 September 9am

Empirical likelihood comparison of t-year absolute risks

Paul Blanche
Department of Biostatistics, University of Copenhagen

In the competing risks setting, which is common in medical research, a key quantity is the t-year absolute risk. It is also often called the cumulative incidence function at time t. In oncology and cardiology, is often estimated with the non-parametric Aalen-Johansen estimator. This estimator handles right-censored data and has desirable large sample properties. Inference for comparing two absolute risks, via a risk difference or a risk ratio, can be done via usual asymptotic normal approximations and the delta-method. However, small sample performance of this approach can be modest. Especially (i) coverage of confidence intervals can be poor and (ii) inference using risk ratios and risk difference can lead to contradictory conclusions, in terms of significant differences. Therefore, we suggest to use empirical likelihood ratio inference as an alternative. This method can be seen as an extension of that of Thomas and Grunkemeier (J Am Stat Assoc, 1975) to the competing risks setting. Interestingly, this approach systematically leads to consistent conclusions when comparing absolute risks via either risk ratios or risk differences, in terms of significance. Simulation results also suggest that small sample inference using this approach can be more accurate. We detail how to compute confidence intervals and p-values using empirical likelihood ratios and provide a software implementation. Novel technical results include formulas and algorithms to compute constrained non-parametric maximum likelihood estimates, from which likelihood ratios and inference procedures are derived. An analysis of clinical data is presented to illustrate the main ideas.

8.5 Contributed - Methods & Theory: Survival Analysis II

Thursday 5 September 9am

Regularised survival analysis of lung cancer patient outcome using radiotherapy dose data

Elizabeth Buckingham-Jeffery
University of Manchester & Highways England

In 2015, over 46,000 new lung cancer cases were identified in the UK. This led to over 35,000 deaths from lung cancer in 2016, making it the UK's most common cause of cancer death. Radiotherapy can be used to destroy cancerous tissues by delivering ionising radiation to cancer cells. The majority of lung-cancer patients treated with curative intent are given radiotherapy and advanced imaging and computing technologies are used to design and deliver these treatment plans. This leads to high-granularity radiotherapy dosing data.Radiation toxicity, for example as a result of radiotherapy damaging healthy tissue, is estimated to be associated with a significant proportion of excess mortality at one year post treatment. We have therefore used survival analysis methods and voxel-level radiotherapy dose data to investigate how radiation delivered to normal lung tissue affects survival in lung cancer patients treated with radiotherapy.We found that due to the high dimensionality of the dosing information it is necessary to use statistical regularisation methods, in particular a fused Bayesian elastic net. Fitting this model to the data allowed us to identify a smooth dose-effect curve. The fitted model was then used to predict how survival could change with (hypothetical) improvements in the delivery of radiotherapy.

This work was completed with Thomas House (University of Manchester), Alan McWilliam (University of Manchester and The Christie NHS Foundation Trust), and Marcel van Herk (University of Manchester and The Christie NHS Foundation Trust).

8.6 Contributed - Methods & Theory: Time Series

Thursday 5 September 9am

Nonparametric clustering for spatio-temporal datasets

Ashwini Venkatasubramaniam[1], Konstantinos Ampountolas[2], Ludger Evers[2]
[1] *The Alan Turing Institute,* [2] *University of Glasgow*

This non-parametric Bayesian approach to clustering for spatio-temporal datasets seeks to identify spatially contiguous clusters that represent distinct temporal patterns. This Bayesian method uses a modified non-sequential distance dependent Chinese restaurant process (ddCRP) to accommodate challenges posed by spatial and network connectivity structures in the graph. The modified ddCRP is also re-defined to enable the number of clusters in the network to be controlled. In addition, a spatio-temporal precision matrix is defined to fully account for within cluster spatial dependencies and the associated temporal patterns. The clustering method uses a Metropolis within Gibbs sampler to explore all possible clustering configurations, as composed of cycles and paths in the network, and infer the relevant model parameters. This novel approach to clustering is able to model multiple dependency structures and determine the number of clusters in a data-driven and computationally efficient manner. This unique method is illustrated by applications to spatio-temporal datasets recorded over a grid-style graph and a map-based network.

8.6 Contributed - Methods & Theory: Time Series

Thursday 5 September 9am

Model averaging of integer-valued autoregressive model with covariates

Jiajing Sun, Yuying Sun, Xinyu Zhang
University of Chinese Academy of Sciences, Academy of Mathematics and Systems Science

This work addresses the model averaging of the integer-valued autoregressive (INAR) process with exogenous covariates. We extended the work of Hansen and Racine (2012) and Gao et al. (2016), and developed leave-subjects-out cross-validation (LsoCV) based model averaging estimator for autocorrelated count time series. We prove that the resulting LsoCV estimator is asymptotically optimal by a criterion equivalent to that used byHansen and Racine (2012). Moreover, the LsoCV estimators are consistent, provided at least one candidate model is not under-fitted. Monte Carlo simulations show that the LsoCV estimators can achieve significant eciency gains over existing model averaging methods.

8.6 Contributed - Methods & Theory: Time Series

Thursday 5 September 9am

Estimation of the effect lag of predictors in modelling temporal datasets with application the effect of lifestyle and air quality on health outcomes in Leeds

Jeanine Houwing-Duistermaat, Haiyan Liu, George Aivaliotis
University of Leeds

The amount of temporal datasets containing relevant information to answer important questions is increasing rapidly. Modelling the relationship between temporal datasets is challenging. Current available models are limited. For example often there is a lag in the effect of predictor on the outcome, e.g. change of lifestyle might not immediately affect your health. Also the effect of historical values of predictors may diminish over time. The available model for this type of question can only deal with one temporal predictor. We propose a functional linear model to predict a temporal response using multiple functional and longitudinal predictors including effect lags of these predictors. Specifically the regression functions in our model are written as the expansion of a basis system (e.g. functional principal components, splines). For a given set of time lags, the coefficients of these fixed basis functions are estimated via optimizing a penalization criterion. The time lags are estimated by simultaneously searching on a prior grid mesh based on minimization of the prediction error. We studied the mathematical properties of the estimated parameters and the accuracy of predicting outcomes based on our model via an extensive simulation study. Finally we applied the methods to data on health (e.g. BMI, COPD), lifestyle (physical exercises) and air quality (e.g. NO2) available at postcode level for the City of Leeds.Our conclusion is that our method performs well and provides insights on the relationships between lifestyle, pollution and health in the City of Leeds. The simulation study showed that the asymptotic properties hold, that the parameter functions are well estimated and that in 70% of the replicates our method identified the correct time lags.

8.7 Contributed - Medical: Estimation and Performance

Thursday 5 September 9am

Estimation of proportions by group retesting

Stephen Walter, Graham Hepworth
McMaster University

To estimate the proportion p of an attribute in a population, sometimes the sampled individuals can be tested in groups. This can yield substantial cost savings over individual testing, especially if p is small. If one desires to increase precision, but it is impractical to sample additional individuals, it may nevertheless be possible to retest groups formed from the individuals within the groups that have tested positive at the first stage. Hepworth and Watson (2017) have described several situations where obtaining additional individuals is indeed impractical or prohibitively expensive. If the initial testing is non-destructive, a re-testing approach may then prove useful. Potential applications include: transmission of diseases by insect vectors; plant or animal disease assessment; prevalence of viruses in mosquito populations; and human blood testing, especially for HIV prevalence. Hepworth and Watson developed an estimator of p for their recommended retesting method, which involves a random regrouping of individuals in the positive groups from the first stage, but because of its analytic complexity, required simulation to examine its variance properties. We have now developed two closed-form analytical expressions for the variance of the second stage estimator, and we compare their performance with the simulated results. Estimates of p from the two stages can be combined for a satisfactory overall estimate. We show that our solutions give acceptable approximations in a reasonable range of circumstances.

References: Hepworth, G. and Watson, R.K. (2017) Revisiting retesting in the estimation of proportions by group testing. Communications in Statistics – Simulation and Computation, 46, 261–274. Walter, S.D., Hildreth, S.W. and Beaty, B.J. (1980) Estimation of infection rates in populations of organisms using pools of variable size. American Journal of Epidemiology, 112, 124–128.

8.7 Contributed - Medical: Estimation and Performance

Thursday 5 September 9am

Survival analysis of cancer patient based on their Genome profile

Khaled Mubarek
Prince Sattam bin Abdulaziz University

Non-small-cell lung cancer (NSCLC) is one of the main sources of death in industrialized nations with an expanding rate around the world. As a result, scientists are now looking for some of the risk factors for lung cancer which can be caused by certain changes in the DNA of lung cells. One way to detect these changes is the copy number alteration (CNA); which is a type of structural variation in the genome . It usually refers to the duplication or deletion of DNA segments larger than 1 kbp. Like other types of genetic variation, some CNAs have been associated with susceptibility or resistance to disease. As a result, CNA can be used to predict the survival of cancer patients. Next-generation sequencing (NGS) technologies produce high-dimensional data that allow a nearly complete evaluation of genetic variation. With the advent of high-dimensional datasets, the following problem has been faced: the number of covariates (in our study 13968) greatly exceeds the number of observations (85). The results of our analysis indicate that we can incorporate the copy number alteration profile to predict the survival time. We investigate a Cox proportional hazards model within a random effects model frame-work using penalized partial likelihood to model the survival time based on lung cancer patients' clinical characteristics as fixed effects and CNA profiles as random effects. We use AIC to estimate σ which parameterizes the covariance variance matrix of the random effect. For the fixed effects the model indicates that age, stageT3, and stageN2 are statistically significant. Finally, comparing the Kaplan-Meier survival curves with model-based average survival function indicates that the model estimation works reasonably well. Also we covered methods for checking the adequacy of a fitted Cox model.

8.7 Contributed - Medical: Estimation and Performance

Thursday 5 September 9am

Assessing performance of survival risk prediction models: A review of traditional and modern methods

David McLernon[1], Ben Van Calster[2], Laure Wynants[3], Maarten van Smeden[2], Ewout Steyerberg[2]
[1] University of Aberdeen, [2] KU Leuven, Leiden University Medical Center, [3] KU Leuven, Maastricht University

Objective: To review methods to evaluate the performance of survival risk prediction models.

Methods & Results: Traditional measures for assessing performance in risk prediction models include the Brier score to indicate overall model performance, the c-statistic for discriminative ability, and calibration models. These methods have been well defined and can readily be applied for the binary outcome setting. Censoring complicates the extension of these methods to survival models, where we may assess performance at specific time points or for the full survival distribution. The Brier score can be extended to the survival setting. The popular extension of the c-statistic for censored data by Harrell discounts pairs that cannot be ordered. However, this approach ignores the study specific censoring distribution on which the population parameters may depend. Uno's C-statistic is a more recent method that models the censoring distribution and uses it to weight the uncensored observations. Calibration-in-the-large is the difference between the mean observed risk and the mean predicted risk, which can be estimated with a model-based approach using the Poisson model. This approach also enables estimation of the calibration slope, i.e. the overall effect of predictors for survival. A relatively new development is to assess the potential impact on making better decisions with decision theoretic measures such as net benefit. This requires the specification of a context dependent decision threshold, or range of plausible thresholds. For illustration of the above methods, we present a case study with internal and external validation of a model predicting recurrence in breast cancer patients following surgery with 2982 women for development and 686 for validation.

Conclusion: Various methods are available to assess overall performance, discrimination, calibration and net benefit, which can be applied for both internal and external validation before using a survival prediction model in clinical practice. Assessing a model's value for making better decisions is essential requiring methods beyond traditional assessments.

8.8 RSS Prize Winners: Best presentations from YSM 2019

Thursday 5 September 9am

Relaxing the constant hazard assumption in a multi-state model in hospital epidemiology

Micki Hill[1], Paul Lambert[12], Michael Crowther[1]
[1] *University of Leicester,* [2] *Karolinska Institutet*

BACKGROUND: Multi-state models are being increasingly used to capture more complex disease pathways. Exponential models have been suggested as an accessible approach to obtain a quick understanding of the data; however, assuming time constant hazards is not always plausible. A flexible spline-based approach has been proposed, where transition hazards can be time-dependent and predictions can still easily be obtained through simulation.

METHODS: Predictions obtained from an exponential model were compared to those obtained via simulation from a Royston-Parmar model with four degrees of freedom. Metrics of interest included: transition probabilities, expected length of stay, attributable mortality (AM) and population attributable fraction (PAF). They (and corresponding confidence intervals) were easily obtained for the spline-based approach using the multistate package in Stata [1]. This work was performed on previously analysed hospital acquired infection (HAI) data [2].

RESULTS: Despite clear deviations from the constant hazards assumption, the empirical estimates of the transition probabilities were approximated reasonably well by the exponential model. However, functions of the transition probabilities, e.g. AM and PAF, were not well approximated and the spline-based approach offered considerable improvements in these cases.

CONCLUSIONS: We conclude that methods and software are readily available for obtaining predictions from multi-state models that do not assume constant hazards. The multistate package in Stata facilitates a range of predictions with confidence intervals via simulation, which can provide a more comprehensive understanding the data.

REFERENCES:[1] Crowther MJ, Lambert PC. Stat Med. 2017;36(29):4719-4742[2] von Cube M, Schumacher M, Wolkewitz M. BMC Medical Research Methodology. 2017;17:111

8.8 RSS Prize Winners: Best presentations from YSM 2019

Thursday 5 September 9am

Calculating Avoidable Deaths for Cancer Patient Survival

Jamie Stokes[1], Mark Rutherford[2]
[1] *University of Oxford,* [2] *University of Leicester*

Background: Differences in cancer survival between population groups are well-documented. The impact of these differences can be determined by estimating the number of deaths that could potentially be avoided or postponed if these differences were to be eliminated.

Methods: Established methods of estimating and modelling relative survival were explored and contrasted in the context of lung cancer patients from the SEER 18 dataset to determine which approach best complements calculation of avoidable mortality. Life table estimates were combined with flexible parametric survival modelling techniques to calculate the number of deaths that could potentially be avoided if survival differences between males and females were to be eliminated. This was explored further, investigating the extent to which age and stage at diagnosis contributes to avoidable mortality.

Results: Flexible parametric models incorporating fractional polynomials for the modelling of age as a continuous variable produced smooth, interpretable and robust estimates of avoidable mortality. Based on an annual cohort of 15971 male lung cancer patients, eliminating survival differences between males and females would lead to postponing approximately 1336 deaths beyond 2 years post diagnosis, representing 11.1% of the deaths predicted to occur in this time. For the same annual cohort, matching the gender distribution in stage at diagnosis would postpone around 378 deaths beyond 2 year post diagnosis, representing 28.3% of the total between-sex avoidable mortality for this period.

Conclusions: Flexible parametric avoidable mortality analysis provides an interpretable method of measuring progress in the reduction of disease survival deficits to inform public health policy, with clear applications in early cancer diagnosis.

8.8 RSS Prize Winners: Best presentations from YSM 2019

Thursday 5 September 9am

In need of some simulation: the Monte Carlo method for sample size calculations

Nick Beckley-Hoelscher, Fiona Reid
KCL

Background: The design stage is one of the most critical aspects of a clinical trial. A major role for the statistician at this stage is estimating the number of participants required. Most widely-used sample size calculations are based on simple mean-based minimal clinically important differences (MCIDs), and on distributional assumptions for which tractable formula for the sample size exist. This framework may break down when calculating sample sizes for non-mean-based outcome measures, complex designs, or situations where there is no tractable formula. Simple, user-written simulations can provide a flexible and easily implementable way of performing sample size calculations in a wide array of situations.

Methods: This presentation will go through the steps required to undertake a simulation sample size calculation, including checks to ensure type 1 error rates are preserved. This will begin with a simple example for illustrative purposes, before moving on to more complex examples.

Results: It is shown how simulations allow for a high degree of user-specification to describe the alternative hypothesis; in particular, describing the MCID in terms of distributional properties other than the mean difference (e.g. percentile-based changes). The examples outlined also show how, provided an alternative hypothesis and method of analysis are specified, the procedure is easily implementable and reproducible.

Conclusion: Simulation-based sample size calculations are a powerful method of calculating the required number of patients for a trial, which can be used as an additional independent method for simple designs, or as a flexible method for more complex designs.

8.9 Contributed - Business, Industry & Finance: Applying statistics to improve industry

Thursday 5 September 9am

Predicting the number of resources required to ensure 97% of the time emergencies are reached in 1 hour.

Laura Thornley
Northern Gas Networks

Northern Gas Networks is a highly regulated business, which means targets are set for the benefits or Customers as well as to ensure high Health and Safety precautions. One of these measures is to ensure that Northern Gas Networks responds to a Gas escape within 1 hour of being informed 97% of the time.Gas escapes vary wildly year to year, as well as having seasonality and links to temperature, meaning time series analysis is used to help support prediction. On top of this, extreme events such as heavy snow or large bursts can cause wide spread events, which require significant number of resources to attend escapes at the same time. The purpose of this analysis was to understand the number of gas escapes that can be expected at any given time (fluctuations daily, weekly and yearly) to allow Northern Gas Networks to plan the number of resources required to attend these escapes.Northern Gas Networks covers the UK area from Berwick upon Tweed down to Doncaster and from Todmorden to Scarborough. Therefore it is important to distribute the resources across the region as it is not possible to travel from on end to another within the 1 hour regulation time frame. As such, spatial analysis is also considered to allow for resources to be distributed in the correct areas and to account for variances in travel time between these regions. A small tolerance is allowed of 3% of escapes to be reached outside of the 1 hour window.Uncertainty analysis has been conducted previously to understand the level of resourcing required to meet the 97% targets and as well as the different levels (98%, 99% and 99.9%). Using these levels of uncertainty allows the business to make informed decisions around resourcing requirements and demonstrate evidence to an auditor that we are managing the network adequately.

8.9 Contributed - Business, Industry & Finance: Applying statistics to improve industry

Thursday 5 September 9am

A new link function to analyze the launch of technological products

Gloria Gheno
Innovative data analysis

In the last decades the most important companies of the technological sector announcethe new products in a spectacular way to intrigue and induce the media to try to predict thenews coming out. In this work I analyze how current and potential customers, based ontheir specific characteristics, are interested in the leaked information, in particularconsidering, for example, the technologic products owned by them and their brand loyalty.Indeed, for a company knowing the target customers interested in launch of newtechnologies is essential to be able to develop products which reflect their needs so as tofully meet their expectations. To study the degree of interest in the advertising campaign, amonth before the launch I ask a sample of people with which probability it assumes that anew product with particular characteristics is launched. Then I compare these results withthe product actually put on the market. I divide the study into two parts, in the first part Ianalyze how the subjects form their own opinions, in the second part I evaluate thegoodness of the forecast. To study the factors which affect these probabilities, I use theBeta regression and a link function which I specifically created. I demonstrate the realgoodness of this new technique by comparing it with others already present in theliterature using specific statistical tests. This work offers companies of the technologysector the opportunity to save time and money by eliminating the costs of development ofunrequired characteristics.

8.9 Contributed - Business, Industry & Finance: Applying statistics to improve industry

Thursday 5 September 9am

Enhancing our understanding of hazard perception in driving

Sritika Chowdhury, Neale Kinnear, Mark Bell
Transport Research Laboratory

Hazard perception is a term used to describe the ability to predict dangerous situations on the road. It is a critical skill that is directly related to safety outcomes. As there is no perfect objective measure of danger within a driving scene, hazard perception is based on subjective appraisal by the driver. IMRA Europe commissioned TRL, the UK's Transport Research Laboratory, and NTU, Nottingham Trent University, to utilise their hazard perception expertise to develop a method to assess danger locations and levels within video footage. One objective of the study was to establish whether physiological measures (of non-conscious arousal) can add value to the validation of danger identification and level. Twenty two drivers with over ten years of driving experience were recruited to view video footage of day-to-day driving from the driver's perspective. Forty one-minute clips and 40 dynamic-stills covered four hazard types: no hazard, precursor not leading to a hazard, precursor leading to a hazard and immediate hazard. Synchronised eye tracking data and physiological measures such as galvanic skin responses (GSR) were recorded while drivers viewed the footage. Drivers were asked to identify the most hazard area of specific scenes and provide a cognitive rating of risk for each video. Finally, drivers were asked to complete a self-reported driving style questionnaire.Analysis involved clustering techniques such as hierarchical clustering and k-means using bootstrapping to cluster eye tracking data and click location; factor analysis of survey data in order to identify specific driving styles in the sample; and exploratory analysis of GSR. A reverse regression model was applied to determine if physiological measures can act as a proxy for determining the level of danger present in a scene and add value to the validation protocol.The data are currently being analysed and the results will be used to inform the development of future research and technology to improve road safety.

PD8 Professional Development: Pre-plenary Overview Meeting

Thursday 5 September 9am

A very short introduction to hypergraph data

Simon Lunagomez Coria
Lancaster University

Network data has become an important resource for both describing phenomena and formulating models in the natural sciences and engineering. Networks have been used to represent systems where units communicate, cooperate or influence one another. Traditionally, relationships encoded in network data involve at most two units. Hypergraphs extend this idea by encoding relationships that involve more than two units (coauthorship of academic papers is a clear example of this). For this talk, we will review different ways to represent a hypergraph as a data structure. We will also discuss some of the modelling approaches that rely on the concept of a hypergraph, either as an observation or as a latent variable.

Keynote 6: Bayesian Categorical Matrix Factorization via Double Feature Allocation

Thursday 5 September 10.10am

Bayesian Categorical Matrix Factorization via Double Feature Allocation

Peter Mueller[1], Yang Ni[2], Yuan Ji[3]
[1] *UT Austin,* [2] *TX A&M,* [3] *U Chicago*

We propose a categorical matrix factorization method to infer latent diseases from electronic health records data. A latent disease is defined as an unknown cause that induces a set of common symptoms for a group of patients. The proposed approach is based on a novel double feature allocation model which simultaneously allocates features to the rows and the columns of a categorical matrix. Using a Bayesian approach, available prior information on known diseases greatly improves identifiability of latent diseases. This includes known diagnoses for patients and known association of diseases with symptoms.For application to large data sets, as they naturally arise in electronic health records, we develop a divide-and-conquer MonteCarlo algorithm, which allows inference for the proposed double feature allocation model, and a wide range of related Bayesian nonparametric mixture models and random subsets.We validate the proposed approach by simulation studies including mis-specified models and comparison with sparse latent factor models. In an application to Chinese electronic health records (EHR) data, we find results that agree with related clinical and medical knowledge.

Bayesian Double Feature Allocation for Phenotyping with Electronic Health RecordsYang Ni, Peter Mueller, Yuan Ji https://arxiv.org/abs/1809.08988

Consensus Monte Carlo for Random Subsets using Shared Anchors,Yang Ni, Yuan Ji, and Peter Mueller https://arxiv.org/abs/[to be posted]

9.1 Medical: The Northern Ireland Clinical Trials Unit (NICTU) and the SANDWICH Trial

Thursday 5 September 11.30am

Stepped Wedge Trial Design and analysis methods for the SANDWICH trial

Clíona McDowell
NICTU

Objective: The objective of this talk is to describe the following:Stepped wedge trial design using the SANDWICH trial as an example.Sample size for the SANDWICH trial.Proposed methods for analysis.

Methods: The app https://clusterrcts.shinyapps.io/rshinyapp/ developed by Karla Hemming was used to calculate the sample size. The aim of the study is to evaluate whether there is a difference in the duration of hours on ventilation before and after exposure to the intervention. There will be censoring i.e. children moving to other units, children not weaned before the unit transitions to the training phase, those who are not weaned at the end of the 20-month trial period, children at the time they have a tracheostomy, those not weaned by 90 days, or children who die. We will use survival analysis and estimate a hazard ratio for the intervention effect. Our survival analysis will estimate the hazard of being extubated and removed from mechanical ventilation.

Conclusions: Cluster randomisation is essential, as the intervention is delivered at the level of the cluster (site) as the individual level components would be susceptible to contamination if individually randomised. The stepped wedge design has been chosen over the conventional parallel cluster design for the following reasons: there are limited number of clusters available to allow detection of the important clinical effect at 90% power; units are more likely to participate in the trial if they are guaranteed their unit will at some point receive the intervention; it would be infeasible and more costly to deliver the intervention simultaneously to units randomised to the intervention in a parallel design; and if the intervention is found to be effective, knowledge translation will be easier as PICUs participating can potentially continue after the trial, maximising the benefits of any effects to the NHS and patients.

9.4 Social Statistics: Old dogs new tricks: respondent centred approach to mixed mode social survey redesign at ONS

Thursday 5 September 11.30am

A respondent centred approach - Questionnaire content redesign for a mixed-mode Labour Market Survey

Vicky Cummings, Alex Nolan
Office for National Statistics

This paper shares insights into how the ONS Social Surveys Transformation Research and Design (R&D) Team has approached the task of redesigning the Labour Force Survey (LFS). This work is being undertaken as part of ONS's Census and Data Collection Transformation Programme which aims to rebalance ONS's data collection activity toward wider, more integrated use of administrative and other non-survey data sources, thereby reducing our reliance on large population and business surveys. The Programme also focuses on developing the capability to move any residual survey data collection online but as part of a mixed-mode design The R&D Team are focusing on redesigning LFS content for online collection and are taking a respondent centered approach to this work. In taking this approach, the R&D Team have aimed to design a survey which is user-centred and optimized for the respondent's chosen mode of completion whilst still meeting the business need to provide high-quality data. This paper outlines the methods employed; including how face to face and telephone interviewers have been integral throughout the development journey. As survey users themselves, the interviewers know where respondents struggle and consulting them has assisted in ensuring existing issues are addressed.This paper explores how the R&D Team has blended traditional cognitive interviewing with user research methods to understand the mental models of respondents around survey topics, survey completion behaviour and usability of survey tools. Through these principles, the full end to end process has been developed and tested with the understanding that giving users the most realistic experience during user testing results in higher quality evidence. By providing practical examples, and drawing on evidence from extensive user testing, this paper demonstrates that, provided respondents understand the concepts, there is scope to design the questions differently to suit the mode and to potentially improve data quality.

9.4 Social Statistics: Old dogs new tricks: respondent centred approach to mixed mode social survey redesign at ONS

Thursday 5 September 11.30am

A respondent centred approach – Social survey materials and respondent engagement strategies for mixed mode collection

Tara McNeill, Natalia Stutter
Office for National Statistics

To successfully achieve the goal of introducing online data collection to its portfolio of household social surveys, ONS is pursuing an innovative research approach, exploring the end-to-end journey of a potential survey respondent. This includes the research and development of respondent materials and engagement strategies which are paramount in getting people to take part independently online. This talk discusses the approach to developing this work and will share key findings.Respondent communications play a pivotal role in multimode data collection, particularly where the approach is online first. Over the last two years ONS has invested heavily in qualitative and quantitative research to explore the impact of respondent materials and mailing strategies on up-take and response rates. The respondent centred approach has demonstrated the potential to achieve a 59% uptake rate for a lightly incentivised push-to-web voluntary social survey.To develop an effective respondent engagement strategy, a blank page approach to materials for household social surveys was taken. This blended traditional qualitative social research methods alongside innovative methods and tools to create a suite of communications based on user insights. The research and design strategy has followed Agile development principles, enabling the iterative development of materials, adapting to changing user needs and maturing government design standards.This research approach has assisted in identifying what works. Through the combined qualitative research and quantitative experiments undertaken as part of the transformation of the Labour Force Survey, valuable insights have emerged about the best mailing strategy for increasing the timeliness of response. Other experiments have explored the extent to which using regionally branded envelopes can break down barriers to opening mail, and also whether attrition on longitudinal surveys can be mitigated through between wave engagement and what medium is most effective. This talk will share the experiences and lessons learnt from developing these materials and engagement strategies.

9.4 Social Statistics: Old dogs new tricks: respondent centred approach to mixed mode social survey redesign at ONS

Thursday 5 September 11.30am

Findings and evidence from large-scale mixed mode testing of transformed social surveys

Colin Beavan-Seymour, Vicky Parker, Joe Herson
Office for National Statistics

ONS is working towards the introduction of online data collection as the primary collection mode for its suite of social surveys, supplemented by traditional face to face and telephone collection. To achieve this, the design for the Labour Force Survey (LFS) has been transformed end-to-end, focussing primarily upon core Labour Market outputs, to produce a shorter and simpler survey instrument. This redesigned LFS is known as the Labour Market Survey (LMS).The LMS is focussed primarily upon providing data for core Labour Market outputs by completely redesigning the content of the survey to make it more relatable and understandable for respondents. This is particularly important for the online self-completion mode. The questionnaire content has a more logical flow and reduces respondent burden through a more efficient routing system. However, the design still has to deliver the same core Labour Market data as the LFS across multiple modes. This has been a challenging, but positive, process.The high profile nature of the LFS and its importance for policy making means that a complete redesign needs to be thoroughly tested, and the impact of changes understood. Since mid-2017, ONS has undertaken a series of iterative quantitative tests; each has had its own distinct research objectives but has also reactively tested changes in the LMS design based upon the successes of each previous test.The testing programme has provided research towards the use of a new sampling frame, incentivisation strategies, materials, online data collection engagement and take up, self-completion questionnaire design, paradata, and national / regional mixed-mode response rates. This talk will discuss how the design of the LMS has evolved since its inception, the testing that has taken place so far, the challenges and opportunities it has presented, and will discuss some preliminary statistical findings which compare LMS and LFS data.

9.5 Methods & Theory: Multi-Parameter Regression Survival Modelling

Thursday 5 September 11.30am

Penalised Variable Selection in Multi-Parameter Regression Survival Modelling

Fatima Jaouimaa[1], Il Do Ha[2], Kevin Burke[1]
[1] *University of Limerick,* [2] *Pukyong National University*

Multi-parameter regression (MPR) modelling refers to the approach whereby covariates are allowed to enter the model through multiple distributional parameters simultaneously. This is in contrast to the standard approaches where covariates enter through a single parameter (e.g., a location parameter). Penalised variable selection has received a significant amount of attention in recent years: methods such as the least absolute shrinkage and selection operator (LASSO), smoothly clipped absolute deviation (SCAD), and adaptive LASSO are used to simultaneously select variables and estimate their regression coefficients. Therefore, in this work, we develop penalised multi-parameter regression methods and investigate their associated performance through simulation studies and real data; as an example, we consider the Weibull model.

9.5 Methods & Theory: Multi-Parameter Regression Survival Modelling

Thursday 5 September 11.30am

Semi-Parametric Multi-Parameter Regression Survival Modelling

Kevin Burke[1], Frank Eriksson[2], Christian Pipper[3]
[1] *University of Limerick, Ireland,* [2] *University of Copenhagen, Denmark,* [3] *LEO Pharma, Denmark*

We consider a log-linear model for survival data, where both the location and scale parameters depend on covariates (i.e., a "Multi-Parameter Regression" [MPR] approach), and the baseline hazard function is completely unspecified. This model provides the fexibility needed to capture many interesting features of survival data at a relatively low cost in model complexity. Estimation procedures are developed and asymptotic properties of the resulting estimators are derived using empirical process theory. Finally, a resampling procedure is developed to estimate the limiting variances of the estimators. A practical application to lung cancer data is illustrated.

9.5 Methods & Theory: Multi-Parameter Regression Survival Modelling

Thursday 5 September 11.30am

A Multi-parameter regression model for interval censored survival data

Defen Peng[1], Gilbert MacKenzie[2], Kevin Burke[3]
[1] *ICVHealth/CHEOS,* [2] *Formerly of Centre of Biostatistics, Dept. of Mathematics & Statistics, University of Limerick, Limerick Ireland,* [3] *University of Limerick,*

We develop flexible multi-parameter regression (MPR) survival models for interval censored survival data arising in longitudinal prospective studies and longitudinal randomised controlled clinical trials. A multi-parameter Weibull regression survival model, which is wholly parametric, and has non-proportional hazards, is the main focus of the paper. We describe the basic model, develop the interval censored likelihood and extend it to include gamma frailty, in order to account for unmeasured covariates, and a dispersion model for the frailty variance, to allow for multiple frailty distributions. We evaluate the models by means of a simulation study and a detailed re-analysis of data from a prospective observational study, the Signal Tandmobiel study. The results demonstrate that The results demonstrate that the multi-parameter regression model with frailty is computationally efficient and provides an excellent fit to the data.

9.6 Communicating & Teaching Statistics: STEM Showcase

Thursday 5 September 11.30am

STEM Showcase

Simon White[1], Rejina Verghis[2], Jennifer Rogers[3]
[1] *University of Cambridge,* [2] *Northern Ireland Clinical Trials Unit,* [3] *University of Oxford*

Statistics, and more importantly the skill of statistical literacy, is fundamental not only to the next generation of STEM (Science, Technology, Engineering & Mathematics) personnel but also to wider society – we must be able to meaningfully critique and discuss the wealth of information used to make policies and run our lives, by governments, business and everyone. As statisticians, we also have a duty to inspire the next generation and to engage the public with mathematics and statistics; to encourage enjoyment of the subject, to enhance and enrich study beyond the curriculum, and to encourage unusual ways of communicating our science.

This session will present examples of engaging statistical activities that have been developed, tested and successfully used in schools and at science festivals to inspire the next generation of statisticians. It will be relevant to statisticians volunteering in schools or with the general public.There will be three talks and a panel discussion.

This first speaker will be Jennifer Rogers, an RSS William Guy Lecturer. This is a prestigious volunteer role which recognises Fellows with a successful track record in undertaking school outreach activities. Jen will discuss her experiences as the post-holder and outline her lecture topic - making life saving decisions in clinical trials: how much evidence do we need?

The second speaker is Rejina Verghis from the Northern Ireland Clinical Trials Unit. Rejina is based in Belfast and has recently registered as a STEM Ambassador. She will outline her motivation for becoming an Ambassador, and the types of activities she hopes to become involved in.

The final speaker will be Simon White from the University of Cambridge, and the RSS Education & Statistical Literacy Committee. Simon will outline several hands-on statistics activities which embed learning important statistical concepts within an approachable (and hopefully enjoyable) context.

Organised by the RSS Education & Statistical Literacy Committee and RSS Young Statisticians' Section

9.9 Methods & Theory: Statistical analysis of relational data

Thursday 5 September 11.30am

Identifying potentially overlapping communities from two-mode networks

Veronica Vinciotti[1], Saverio Ranciati[2], Ernst Wit[3]
[1] *Brunel University London,* [2] *University of Bologna,* [3] *Universita' Svizzera Italiana, , , , , , ,*

Actor-event data are common in sociological settings, whereby one registers the pattern of attendance of a group of social actors to a number of events. These data are often transformed to actor-actor data in order to be further analysed by network models, such as stochastic block models. This transformation and such analyses lead to a natural loss of information, particularly when one is interested in identifying, possibly overlapping, subgroups or communities of actors on the basis of their attendances to events. In this talk, we will present an actor-event model for overlapping communities and develop a Bayesian procedure for inference. We present a real application on a terrorist network, whose attendance to events was monitored during a period of time.

9.9 Methods & Theory: Statistical analysis of relational data

Thursday 5 September 11.30am

A network approach to votes exchange in the Eurovision Song Contest

Silvia D'Angelo[1], Thomas Brendan Murphy[1], Marco Alfò[2]
[1] *University College Dublin,* [2] *Sapienza, University of Rome*

The Eurovision Song Contest is a popular TV singing competition held annually among country members of the European Broadcasting Union. In this competition, each member can be both contestant and jury, as it can participate with a song and/or vote for other countries' tunes. During the years, the voting system has repeatedly been accused of being biased by tactical voting; votes would represent strategic interests rather than actual musical preferences of the voting countries. In this work, we develop a latent space model to investigate the presence of a latent structure underlying the exchange of votes. Focusing on the period from 1998 to 2015, we represent the vote exchange as a multivariate network: each edition is a network, where countries are the nodes and two countries are linked by an edge if one voted for the other. The different networks are taken to be independent replicates of a conditional Bernoulli distribution, with success probability specified as a function of a common latent space capturing the overall relationships among the countries. Proximity denotes similarity, and countries close in the latent space are more likely to exchange votes. If the exchange of votes depends on the similarity between countries, the quality of the competing songs might not be a relevant factor in the determination of the voting preferences, and this would suggest the presence of some bias. A Bayesian hierarchical modelling approach is employed to estimate the parameters, where the probability of a connection between any two countries is a function of their distance in the latent space, network-specific parameters and edge-specific covariates.

9.9 Methods & Theory: Statistical analysis of relational data

Thursday 5 September 11.30am

Latent Space Representations of Hypergraphs

Kathryn Turnbull[1], Simon Lunagomez[1], Christopher Nemeth[1], Edoardo Airoldi[2]
[1] *Lancaster University,* [2] *Fox School of Business, Temple University*

Relational data describing interactions among a population arise in a multitude of disciplines, including systems biology, neuroscience and marketing. There exists a broad literature concerned of analysis of such data when the interactions are assumed to be pairwise. However, in many real-world applications the interactions may instead occur between several members of a population and, in this case, the data are more appropriately represented by a hypergraph. As an example, consider a coauthorship network where an interaction indicates which academics have contributed to a paper. It is common for a group of authors larger than two to write an article jointly, which corresponds to a hyperedge. The literature on statistical analysis of hypergraphs is relatively underdeveloped, and in this talk we introduce a model which extends the latent space approach of Hoff et al (2002) for graphs to the hypergraph setting. In this framework, the nodes of the hypergraph are assumed to lie in a low-dimensional space and the hyperedges are modeled as a function of the latent coordinates. Using a toolkit from stochastic geometry, we develop a computationally efficient model with a convenient likelihood. Furthermore, we explore and analyse the properties of this model, and use the latent space to perform predictive inference.

This is join work with Simon Lunagomez, Christopher Nemeth and Edoardo Airoldi.

9.9 Methods & Theory: Statistical analysis of relational data

Thursday 5 September 11.30am

Methods of analysing and comparing networks

Sofia Olhede[1], Patrick Wolfe[2]
[1] *EPFL & UCL,* [2] *Purdue University*

Networks have become a key data analysis tool. They are a simple method of characterising dependence between nodes or actors. Understanding the difference between two networks is also challenging unless they share nodes and are of the same size. We shall discuss how we may compare networks and also consider the regime where more than one network is observed.

10.1 Contributed - Medical: Meta-analysis

Thursday 5 September 2pm

On the comparison of alternative models in dose-response meta-analysis using summarized data

Nicola Orsini
Department of Public Health Sciences, Karolinska Institutet

Objectives: A linear-mixed effects model for the synthesis of multiple tables of summarized dose-response data has been recently proposed. The main advantage is to include studies contrasting just two doses into spline analysis. Our aim is to evaluate the ability of the Akaike's information criterion (AIC) to suggest the underlying dose–response relationship.

Methods: Ten individual studies of 1,000 persons each were simulated under the assumption of either a linear (Shape 1) or non-linear (Shape 2) relationship between a quantitative dose and the mean outcome. The dose followed a positive right-skewed distribution. Tables of summarized data (mean dose, mean difference about referent, sample size, standard deviation) were generated upon categorization of the dose into two-quantiles. Every simulated dose-response meta-analysis was analyzed with a linear-mixed effects model using two commonly used strategies: linear and restricted cubic splines with 3 knots. Accuracy of the AIC was examined by calculating the proportion of times in 1,000 simulations the Shape 1 and Shape 2 were correctly identified by choosing the lowest AIC among the modelling strategies. We next explored how this accuracy could improve by randomly categorizing the dose using two or three quantiles.

Results: Under Shape 1, the proportion of times the lowest AIC correctly indicated the linear function was 99%. This proportion did not change using a mix of two/three quantiles to create tables of summarized data. Under Shape 2, the proportion of times the lowest AIC correctly indicated non-linearity was 2%. Using a mix of two/three quantiles raised the accuracy to 98%.

Conclusions: The simulations of dose-response meta-analysis based on summarized data using linear mixed-effects models showed a good and stable accuracy of the AIC to correctly identify underlying linear relationships with just two quantiles of the dose. Regarding non-linear relationships, good accuracy of the AIC can be achieved with a random mix of two and three quantiles.

10.1 Contributed - Medical: Meta-analysis

Thursday 5 September 2pm

An extended mixed-effects framework for meta-analysis

Francesco Sera, Antonio Gasparrini
London School of Hygiene and Tropical Medicine

Objectives: Meta-analysis has become a standard method to summarize evidence in clinical research. Standard applications commonly consider a single effect size estimated from independent studies. However, extensions to deal with more complex meta-analytical problems have been presented (e.g. multivariate, multilevel, and longitudinal meta-analysis). All these extensions can be described as cases where non-independence among observation within studies creates more complex correlation structures. The aim of this contribution is to develop a unified framework for meta-analysis based on linear mixed-effects (LME) models, where the correlation between effect sizes is modelled through a flexible random-effects structure.

Methods: We derive a general analytic formulation that includes, as special cases, all the specific models mentioned above. We have defined (restricted) maximum likelihood (ML and REML) estimators, with efficient computational strategies based on profiled methods that alternate (restricted) iterative generalized least squares and Newton-Raphson procedures. Inference is based on asymptotic multivariate normal distribution of (RE)ML estimates. The analytic framework and the inferential procedures are implemented in the new R package mixmeta.

Results: The modelling framework will be illustrated using three case studies. In particular, we will present an application of multivariate meta-analysis on 24 trials that compare four alternative interventions to promote smoking cessation. In a second example we consider a multilevel meta-analysis with 38 observations within 20 trials, each evaluating the association between thrombolytic therapy and short-term mortality risks in multiple sub-groups of treatment delay after a myocardial infarction. Finally, we illustrate an application of longitudinal meta-analysis using data on 17 randomized controlled trials comparing treatments of malignant gliomas at 6, 12, 18, and 24 months since the start of the treatment

Conclusions: The definition of a unified framework for meta-analysis, complemented with the implementation in a freely-available and fully documented software, will provide researchers with a flexible tool for addressing non-standard pooling problems.

10.1 Contributed - Medical: Meta-analysis

Thursday 5 September 2pm

Synthesis of individual and aggregate level data using multilevel network meta-regression: extension to general likelihoods

David Phillippo[1], Sofia Dias[2], Tony Ades[1], Nicky Welton[1]
[1] *University of Bristol,* [2] *University of York*

Standard network meta-analysis (NMA) and indirect comparisons combine aggregate data (AgD) from multiple studies on treatments of interest, assuming that any effect modifiers are balanced across populations. Individual patient data (IPD) meta-regression can relax this assumption, but in many cases IPD are only available in a subset of studies. Multilevel Network Meta-Regression (ML-NMR) is a recently-proposed method which extends the IPD meta-regression framework by integrating the individual-level model over AgD covariate distributions to incorporate data from IPD and AgD studies and avoid aggregation bias. However, ML-NMR requires the aggregate-level likelihood to have a known closed form. Most notably, this precludes the application of ML-NMR to synthesis of time-to-event outcomes, which make up the large majority of population adjustment analyses to date.We extend ML-NMR to handle individual-level likelihoods of general form, illustrating with two examples – a real network of plaque psoriasis treatments with ordered categorical outcomes, and a simulated comparison of time-to-event outcomes.We show how the individual-level likelihood function conditional on the covariates is integrated over the covariate distributions in each AgD study to obtain the respective marginal likelihood contributions. Quasi-Monte Carlo numerical integration is used, making application general.Joint synthesis of ordered categorical outcomes lead to increased precision compared to separate models. ML-NMR achieved better fit than a random effects NMA, uncertainty was substantially reduced, and the model was more interpretable. For the simulated survival data, ML-NMR agreed closely with the known truth, with little loss of precision from a full IPD analysis.ML-NMR is a flexible and general method for synthesising evidence from mixtures of individual and aggregate level data in networks of all sizes. Extension to general likelihoods, including for survival outcomes, greatly increases the applicability of the method. Decision making is aided by the production of effect estimates relevant to the decision target population.

10.2 Contributed - Official & Public Policy: Business and migration

Thursday 5 September 2pm

Design strategy for data collection in business surveys to reduce response bias

Dominic Brown, Gary Brown
The Office for National Statistics

In the current era of falling response rates and reduced budgets, National Statistical Institutes (NSIs) are compelled to innovate in data collection strategies. Instead of a single static approach, designs that are adapted to different populations, and respond to different conditions, are increasingly being considered.The UK Office for National Statistics (ONS) has survey-specific targets for response rates for the 80+ establishment surveys it conducts annually. During the data collection phase of the survey cycle cases are prioritised for data collection, known in the UK as ‘response chasing’. The research project reported in this talk examines the optimal response chasing strategy – in terms of cost and quality – for the ONS as well as exploring some innovative possibilities and assess the implications for production of official statistics.

10.2 Contributed - Official & Public Policy: Business and migration

Thursday 5 September 2pm

Using linked administrative data to better understand the activity and contributions of international migrants within the UK

Megan Bowers, Nicola Rogers, Becca Briggs
ONS

International migrants enter and leave the UK for a variety of reasons, stay for different lengths of time and interact with society and the economy in different ways. As part of the cross-Government Statistical Service migration statistics transformation programme, ONS are making use of a greater range of data to improve our understanding of these complex issues and provide more detailed and relevant statistics to our users. We also need to provide clear insights into how different groups impact on our workforce, communities and public services such as the NHS and schools. We recently published a research engagement report which set out our ambition to produce new insights on the impact of migration and showed progress towards our new approach for producing international migration flows and population statistics using administrative data. A series of in-depth case studies were presented demonstrating how we can use administrative data to identify "activity" for the migrant population. Activity can be defined as an individual interacting with a system, for example attending a hospital or claiming a benefit. We discovered differences in activity depending on migrants' nationality, finding evidence that EU nationals register more quickly for a National Insurance Number than non-EU. We also found that around 70% of international students departed for 12 months or more at the end of their study in England and Wales, while a quarter extended their UK stay.The case studies put a spotlight on what linking together specific data sources – such as immigration, education, health and tax records – can tell us about international migration. This presentation will cover our latest findings, why they matter, our plans to produce further insights and an opportunity to gather user feedback on the transformation programme. It will provide further context to the RSS conference session on transforming population, migration and social statistics.

10.2 Contributed - Official & Public Policy: Business and migration

Thursday 5 September 2pm

How are services traded internationally? UK developments measuring the trade in services by modes of supply.

Daniel Cheung, Dean Scott
Office for National Statistics

With the services industry representing approximately 80% of UK economic activity, following the EU referendum there has been a growing need for trade in services data by modes of supply; detailing the methods in which services are traded globally. Through international collaboration with organisations such as the US Bureau of Economic Analysis and with other government departments, the Office for National Statistics (ONS) has been leading the way with this research.The data allows for a much more detailed understanding of how the UK service sector operates globally, assisting trade negotiations by highlighting strengths and weaknesses within the sector and allowing negotiations to be targeted for best impact. Analysts will benefit by having a greater understanding of the globalisation of the UK's services sector.There are four modes of supply as defined by the World Trade Organisations General Agreement on Trade in Services (GATS): trade conducted remotely, trade provided by a commercial subsidiary present within the host country, trade conducted by the supplier traveling abroad to provide the service, or the customer traveling to the supplier's home country to obtain the service. ONS piloted new mode of supply questions as part of the International Trade in Services survey: the UK's biggest source of trade in services data. Following this pilot, the new questions were included as part of the full annual survey. This presentation will detail these recent developments and present the latest findings, with ONS being one of the first countries to publish results on modes of supply.

10.3 Contributed - Applications of Statistics: Applications 2

Thursday 5 September 2pm

Probability reasoning in judicial fact-finding

Ian Hunt[1], The Honourable Mr Justice Mostyn[2]
[1] *Monash University,* [2] *Royal Courts of Justice*

We argue that the laws of probability help promote coherent fact-finding ("on-the-balance-of-probability") and can avoid logical contradictions. We assume that the laws of probability hold, that Bayes' formula is valid and that probability can be interpreted as subjective degrees of belief. Our argument is essentially that an "elementary probabilistic model of degrees of belief often contains just the right balance of accuracy and simplicity to enable us to command a clear view of the issues and see where we [are or could be] going wrong" (Horwich, 1993).For example, assume that there are three mutually exclusive and exhaustive explanations for something that happened and that the judge reckons that the probability of each explanation being true is less than 0.5. Also assume that on pain of incoherence the judge ensures that the sum of her subjective probabilities is one. In this case, no legal facts can be found on-the-balance-of-probabilities; to find otherwise would imply a contradiction in terms of the laws of probability (in particular that the three probabilities must sum to one) or require a post-hoc fix to the originally reckoned probabilities. Worse than a mere contradiction or a fix-up of the odds, a finding of fact for an event with probability less than 0.5 risks serious injustice.We analyse real cases in which judges apply similar reasoning and a related appeal decision which admonishes judges to avoid using the laws of probability in findings of fact (it goes as far as suggesting that a reference to the probability of a past event is pseudo-mathematics). We respond to the criticisms launched in the appeal judgment and set out a positive case for using probability reasoning in judicial fact-finding. And we argue that probabilistic reasoning enlightens, in a therapeutic sense, legal principles related to inherent probability, the Binary Method and the blue bus paradox (calibrating subjective probabilities to statistical frequencies).

10.3 Contributed - Applications of Statistics: Applications 2

Thursday 5 September 2pm

Bayesian networks and chain event graphs as decision making tools in forensic science

Gail Robertson, Amy Wilson
University of Edinburgh

Bayes' theorem and likelihood ratios are used in forensic statistics to compare evidence supporting different propositions put forward during court proceedings. There is widespread interest among forensic scientists in using Bayesian network models to evaluate the extent to which scientific evidence supports hypotheses proposed by the prosecution and defence. Bayesian networks are primarily used to compare support for source-level propositions, e.g. those concerned with determining the source of samples found at crime scenes such as hair, fibres, and DNA. While comparing source-level propositions is useful, propositions which refer to criminal activities (i.e. those concerned with understanding how a sample came to be at the crime scene) are of more interest to courts. Less work has been done on developing probabilistic methods to assess activity-level propositions, hence finding a method of evaluating evidence for these types of propositions would benefit practitioners. Chain event graphs have been suggested as a decision making tool to assess the extent to which evidence supports event timelines proposed by the prosecution and defence, and may be more appropriate for assessing activity-level propositions. In this study we used Bayesian networks and chain event graphs to combine different types of evidence supporting activity-level propositions from a real-world drug trafficking case. We compared the use of Bayesian networks and chain event graphs in evaluating evidence from activity-level propositions associated with the case and developed a framework for these to be used by practitioners with non-statistical backgrounds. We found that chain event graphs were better suited at evaluating evidence from complex event timelines put forward by the prosecution and defence.

10.3 Contributed - Applications of Statistics: Applications 2

Thursday 5 September 2pm

Judging a book by its cover - How much of REF 'research quality' is really 'journal reputation'?

David Selby, David Firth
University of Warwick

The Research Excellence Framework (REF) is a periodic UK-wide assessment of the quality of published research in universities. The most recent REF was in 2014, and the next will be in 2021. The published results of REF2014 include a categorical 'quality profile' for each unit of assessment (typically a university department), reporting what percentage of the unit's REF-submitted research outputs were assessed as being at each of four quality levels (labelled 4*, 3*, 2* and 1*). Also in the public domain are the original submissions made to REF2014, which include—for each unit of assessment—publication details of the REF-submitted research outputs.In this study we address the question: to what extent can a REF quality profile for research outputs be attributed to the journals in which (most of) those outputs were published?The data are the published submissions and results from REF2014. The main statistical challenge comes from the fact that REF quality profiles are available only at the aggregated level of whole units of assessment: the REF panel's assessment of each individual research output is not made public. Our research question is thus an 'ecological inference' problem, which demands special care in model formulation and methodology. The analysis is based on logit models in which journal-specific parameters are regularized via prior 'pseudo-data'. We develop a lack-of-fit measure for the extent to which REF scores appear to depend on publication venues rather than research quality or institution-level differences. Results are presented for several research fields.

10.4 Contributed - Social Statistics: Response Data

Thursday 5 September 2pm

Administrative Data and it's use to combat survey Non-Response

Matthew Moore, Alasdair Rutherford
University of Stirling

The issues arising due to non-response are problematic for many surveys. With an international trend of dropping survey response rates (Zhang et al, 2013). This has led to more researchers asking the extent to which can non-response can cause issues with analysis, and what can be done to combat this non-response. There are a few ways to approach this problem which can be grouped into either fieldwork or post-fieldwork techniques. Non-response weighting is a common method used, aiming to reduce potential non-response bias. However, the construction of these weights often rely upon untested assumptions and low quality and quantity of data (Collins et al 2001). Administrative data can provide new opportunities in survey design including the construction of non-response weights allowing the production of better survey-based estimates. In this paper we explore the potential of using administrative data from the sampling frame of a large social survey to model non-response. The data used for this paper is part of the Healthy Aging in Scotland (HAGIS) Survey. The HAGIS pilot wave collected a sample of approximately 1000 individuals aged over 50 in the Scottish mainland. HAGIS utilised administrative data in the sampling with screening being provided to ensure households with at least one eligible respondent (over 50) was selected. Administrative data was also collected available for the full sample, and not just respondents to the HAGIS pilot wave. The administrative data consists of individual demographics, hospital admissions, and prescription data held by NHS Scotland. Survey response is modelled as a function of explanatory variables from the administrative data in order to examine whether there is systematic demographic and health-based non-response in the survey. We further examine whether this non-response is causing significant biases and then how health data might predict non-response. Finally, we discuss how these models could be used in improving conventional weighting techniques to combat non-response in survey design.

10.4 Contributed - Social Statistics: Response Data

Thursday 5 September 2pm

On reproducibility of hypothesis tests based on randomised response data

Fatimah Alghamdi
Durham University

Randomised response techniques (RRT) are frequently used when data on possiblysensitive information is being collected by using a survey. There are many different RRT methods and strategies which seek to know the truth from the respondents with more efficiency and more privacy and without embarrassment beside its ability to decrease the bias which can happen by wrong answers; the first techniques was presented by Warner (1965), then Greenberg model (1979) who has modified some properties in Warner.A question of interest in hypothesis test scenarios is the reproducibility of the results: if the test was repeated, would it lead to the same conclusion with regard to rejection of the null hypothesis? We address this question for Warner's and Greenberg method. We use nonparametric predictive inference, a frequentist approach based on only few assumptions, to derive lower and upper probabilities for test reproducibility.This poses the challenging question of finding another measurement, which is called the measurement of the lower reproducibility probability (MRP), to compare between the Warner and Greenberg test.Greenberg model is more efficient than Warner model and the measurement MRP and the area of non-rejection of the measurement (AUMRP) of Greenberg test are higher as well.The work will continue to explore more useful results and real applications about (RRT) and it can be developed in another direction.

Reference[1] Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association 60(309), 63–69.[2] Greenberg, B. G., Abul-Ela, A. L. A., Simmons, W. R., Horvitz, D. G. (1969). The unrelated question randomized response model: Theoretical framework. Journal of the American Statistical Association, 64(326), 520-539.

10.4 Contributed - Social Statistics: Response Data

Thursday 5 September 2pm

Response rates, fieldwork and survey quality: Does reissuing reduce non-response bias?

Linda Hutcheson
Ipsos MORI

Traditionally, response rates have been used as a proxy of survey quality. High response rates are assumed to indicate good quality survey data with less potential for non-response bias. However, over the last twenty years, response rates have been declining. Most major face-to-face random probability surveys have expended more and more fieldwork effort and cost by increasingly reissuing non-responding sample at first issue to other interviewers to try to maintain response rates.But does this reduce non-response bias? This study examines the impact that reissuing has on key survey estimates from two of Scotland's most important population surveys: the Scottish Household Survey and the Scottish Crime and Justice Survey. In these surveys, reissuing increases the response rate by around 8-10%, from 55-60% to around 63-68%. The impact on a range of measures is examined by comparing the published estimates from the weighted full survey sample with estimates from first issue interviews only, weighted as if they were the final achieved sample. As well as examining the impact at the overall national level, impact among key sub-groups is assessed, and results from two waves of each survey are compared to examine whether impact is consistent across different waves.Overall, the results show that the impact of increasing the response rate through reissuing on national estimates is modest. Among sub-group estimates, while the impact in absolute terms is larger, this is because these estimates themselves are less precise as they are based on smaller sample sizes. Reissuing some forms of unproductive case may have more impact on estimates than others.

10.5 Contributed - Methods & Theory: High-Dimensional Data

Thursday 5 September 2pm

A novel method for analysis of high dimensional data

Reza Drikvandi
Manchester Metropolitan University

High dimensional data are rapidly growing in many domains, for example, in microarray gene expression studies, fMRI data analysis, large-scale healthcare analytics, text/image analysis, natural language processing and astronomy, to name but a few. In the last two decades regularisation approaches have become the methods of choice for analysing high dimensional data. However, obtaining accurate estimates and predictions as well as reliable statistical inference remains a major challenge in high dimensional situations. In this talk, we introduce a novel method to overcome this challenge in high dimensional linear regression models. The proposed method enjoys from an effective combination of the regularisation methods such as lasso and the dimensionality reduction techniques such as PCA and sparse PCA. The sparse PCA is scalable and reduces the computations dramatically, which is very attractive for big data analysis. The proposed method is evaluated theoretically and practically using simulated and real data analysis, and its performance is compared with existing regularaisation methods. We show, via theoretical and simulation investigations, that the proposed method produces accurate parameter estimates and predictions, and also enables us to conduct statistical inference on regression coefficients. Finally, we briefly discuss valid post-selection inference in high dimensional data using the proposed methodology.

10.5 Contributed - Methods & Theory: High-Dimensional Data

Thursday 5 September 2pm

Bias reduction methods for binary regression with misclassified responses

Euloge Clovis Kenne Pagui, Alessandra Salvan, Nicola Sartori
Department of Statistics - University of Padoua

Methods for binary regression models that ignore misclassification of the response may lead to a substantial bias on the maximum likelihood estimators of the regression parameters. With known misclassification probabilities, a flexible maximum likelihood framework which allows for misclassification of the responses is possible and has been applied in different contexts. Unfortunately, with moderate sample information or high misclassification probabilities, the bias of the maximum likelihood estimator can still be large and may result in misleading inference. As a solution to this problem, we propose the use of the adjusted score functions for mean and median bias reduction proposed by Firth (1993) and Kenne Pagui et al. (2017), respectively. Both approaches, consist of adding a suitable adjustment term to the score function and then solving the resulting adjusted score equation. The former approach produces a second-order mean bias reduced estimator while the latter gives a third-order median bias reduced estimator. The two methods do not require finiteness of the maximum likelihood estimate and are effective in preventing infinite estimates. For the implementation of mean and median bias reduction, a simple modification of existing software for generalized linear models can be used. Simulation studies show that mean and median bias reduced estimators have better frequentist properties than standard maximum likelihood.

10.5 Contributed - Methods & Theory: High-Dimensional Data

Thursday 5 September 2pm

Refining clustered standard errors with few clusters

Gianmaria Niccodemi, Tom Wansbeek, Rob Alessie, Viola Angelini, Jochen Mierau
University of Groningen

White cluster-robust standard errors, based on OLS residuals, are typically underestimated in case of few clusters and positive correlation of the error term within cluster. This often implies misleading inference, especially on constant within cluster regressors. Common methods to estimate unbiased clustered standard errors with few clusters include the so-called CR2VE, also known as BRL (bias reduced linearization), and CR3VE, both based on transformed OLS residuals. An alternative method is based on estimating the variance components to directly control for within-cluster correlation of the error term, if the covariance matrix of the error term is known to have a certain within-cluster parametrization (e.g., random effects). In this paper we present three refinements for these methods. First, we introduce CR3VE-λ, an adaptation for CR3VE that takes the different cluster sizes into account to make the computed standard errors closer to unbiasedness. Second, we introduce formulas to compute equivalent CR2VE and CR3VE that exploit lower order matrices and that are therefore computationally more efficient. Third, we show how the standard Swamy-Arora random effects model, implemented in most statistical packages (e.g., Stata), can perform poorly with few clusters as its estimate of the cluster effect can be negative. This problem can be mitigated by employing a different random effects estimator. We illustrate our refinements by several Monte Carlo simulations that show that our CR3VE-λ performs better than CR3VE in case of highly unbalanced clusters, that our equivalent CR2VE and CR3VE are computationally much faster as the cluster sizes increase, and that the simple alternative to Swamy-Arora random effects model may be adequate for the computation of clustered standard errors with few clusters, regardless the nature of the covariance matrix of the error term.

10.6 Contributed - Data Science: Rage Against the Machine Learning

Thursday 5 September 2pm

Treating missing data with machine learning

Vinayak Anand Kumar, Vahe Nafilyan, Alex Milroy
Office for National Statistics

Missing data can be problematic as they reduce the reliability and accuracy of statistics. Imputation generates values and/ or units to produce a dataset that is more representative of the population and concept of interest. Ideally, imputation methods should be tailored to specific problems, including the nature of missingness. . Unfortunately, in practice imputation models are not always empirically tested due to the large volume of data or timeliness constraints. The Methodology Division of the Office for National Statistics (ONS) in collaboration with the Data Science Campus investigate the use of supervised Machine Learning (ML) techniques to carry out imputation.. We test several state-of-the-art ML algorithms such as XGBoost and Generative Adversarial Networks (GANs) to impute missing values and compare this to the standard approach. We also use ML techniques to simulate missingness and create a test dataset that we then use to validate the imputation methods. The presentation will cover the key concepts behind XGBoost and GANs and the findings from this program of work.

10.6 Contributed - Data Science: Rage Against the Machine Learning

Thursday 5 September 2pm

Predicting Human States with Multimodal data utilising Machine Learning Methodologies

Daniel Moore[1], Christine Spencer[1], Lucy Rutherford[2], Gawain Morison[2], Gary McKeown[1], Lisa McFetridge[1], Ben Bland[2]
[1] *Queen's University Belfast,* [2] *Sensum*

The future of in-car sensing technology is facing a radical change over the coming decade, largely driven by the need for increased driver safety. In this endeavour various approaches have been explored, from self-driving vehicles to active collision-avoidance systems. However, the logic by which these technologies should best act to remove the driver's influence from the car is not well-defined. Therefore, our objective is to identify the driver's latent stress to potentially facilitate the automatic handover of driver control systems, and automate the change of in-cabin environment to reduce this latent stress. Such a 'reactive-cockpit' could be activated through various systems, including the vehicle's infotainment devices, HVAC (climate) control, speed-limiter and assisted-driving technologies.This work explores a multimodal approach to detecting driver stress. Here, human physiology, facial expression and voice are analysed together to enhance the prediction accuracy in a method that is more robust than using one data type alone. This work details the end-to-end steps of developing a real-time machine-learning based technique for predicting stress, and the problems and caveats associated with such a project. Here, data is collected using a driving simulator and real-world driving tasks, using protocols designed to stimulate the ground-truth state. Blood pressure and self-report data are used to confirm the protocols' effectiveness. The data collected is then annotated on a continuous scale and used as the training dataset. Features are derived from the raw signals, such as Heart Rate Variability and Blink Rate, as predictors on which the model is trained. The results indicate that this approach is able to effectively detect driver stress as confirmed by the self-report data and blood pressure values. Furthermore, such a model may be used to identify prolonged periods of driver stress, which can act as a trigger-point for the car's safety and environmental control systems, which in turn may increase driver safety and comfort.

10.6 Contributed - Data Science: Rage Against the Machine Learning

Thursday 5 September 2pm

Rage Against The Machine - The limits of maching learning for automatic product classification

Jens Mehrhoff
European Commission

The talk contains an assessment of applying supervised learning classification algorithms to automatically map scanner data to the official classification of sub-indices for consumer price statistics. The main test objectives are summarised as follows:compare the family of probabilistic supervised machine learning classification methods;use more or less detailed product descriptions and identifiers as feature-generating variables; andconsider different percentages of the data set as training data - from smaller and larger data sets.A major challenge in this respect is textual feature engineering. The product descriptions are not natural text but use specific vocabularies and rely on different kinds of shorthand. To this end, all possible three-character strings (trigrams) are exploited.The performance measures derived are precision and recall as well as the execution time. Particular emphasis will be given to the interpretation of the results in terms of usability in the production process, i.e. guidance on what works, what does not, and why.

10.7 Contributed - Medical: Hospital Outcomes

Thursday 5 September 2pm

Using funnel plots and CUSUM techniques to monitor hospital-standardised mortality

Chris Mainey
University Hospitals Birmingham NHS Foundation Trust

Patient mortality and its association with healthcare is a contentious topic, with a strong academic history, and various entrenched views. Despite differing opinions, hospitals and regulatory bodies routinely measure and monitor mortality rates, regarding them as 'smoke alarms' related to quality of care. Many of the techniques used for monitoring were described in an influential RSS paper by Spiegelhalter et al. (2012), and this talk describes them in practice in the NHS.The risk of death varies between patients according to many factors and these factors must be taken into account when monitoring mortality. Two of the most common indicators: the Summary Hospital-level Mortality Index (SHMI) and the Hospital Standardised Mortality Ratio (HSMR) are calculated as indirectly-standardised ratios of observed and 'expected' deaths. They vary in their predictors and definitions, but both use logistic regression to predict the risk of death per patient, and the sum of the risk scores form counts of 'expected' deaths at organisations. These methods are commonly presented as 'funnel plots,' and use control limits to identify outlier organisations. These limits are based on the Poisson distribution and suffer from overdispersion, and this talk will discuss adjustment techniques. Funnel plots are only suited to cross-sectional analysis and other techniques are need to monitor over time. This talk will discuss risk-adjusted CUSUM techniques, as used by NHS regulator the Care Quality Commission (CQC) and Imperial College's mortality monitoring system. It explains how they are calculated, the differences between techniques, and how they are used in practice.Two alternative approaches will be discussed: CUSUMs based on aggregated data using z-scores, and patient-level log-likelihood charts. These charts have different processes for setting trigger thresholds and involving simulation or approximation techniques. Examples of techniques will be shown using R, and shortcomings such as the effects of overdispersion and false positives, will be discussed.

10.7 Contributed - Medical: Hospital Outcomes

Thursday 5 September 2pm

So, you've been added to the lung transplant list, now what? An analysis of patient outcomes from listing for a lung transplant

Rachel Hogg[1], Antonios Kourliouros[2], Jenny Mehew[1], Mohamed Al-Aloul[3], Martin Carby[4], James Lordan[5], Richard Thompson[6], Steven Tsui[7], Jasvir Parmar[7],
[1] NHS Blood and Transplant, [2] Royal Brompton Hospital, [3] Wythenshawe Hospital, [4] Harefield Hospital, [5] Freeman Hospital, [6] Queen Elizabeth Hospital Birmingham, [7] Royal Papworth Hospital

Objective: When listed for a lung transplant, patients have several outcomes on the list; transplant, death, or removal. Demand for lung transplantation vastly exceeds the availability of donor organs, translating into long waiting times and high waiting list mortality. The purpose of this work was to identify risk factors associated with the different patient outcomes on the list, with results being used as a tool to help patients understand their individual risks of each outcome (in collaboration with the Winton Centre).

Methods: Adults first registered for lung transplantation in the UK between 1 January 2004 and 31 March 2014 were analysed. Cox regression models were developed for each outcome to identify risk factors and their associated impact. Due to the presence of competing risks, Fine and Gray models were used to provide parameter estimates and risk-adjusted cumulative incidence functions for each event.

Results: At 2 years post-registration, 58% of the 2213 patients were transplanted and 22% had died on the list. A range of factors were found to impact each outcome from the Cox analyses. Chance of transplant varied by disease group and centre ($p<0.001$). Taller patients ($p<0.001$) had an increased chance of transplant. Age had a non-linear effect on death on the list with recipients aged 42-56 having the highest chance of death.

Conclusion: Patients on the lung transplant list are complex with a range of factors influencing their chance of a transplant or other outcome. This work identified these factors, informing the clinical community and the future direction of allocation policy, allowing for better patient care. The work with the Winton Centre will improve information available to patients regarding their potential outcomes.

10.7 Contributed - Medical: Hospital Outcomes

Thursday 5 September 2pm

"Tell us what the data say": Understanding and misunderstanding control charts for monitoring hospital outcomes

Peter Martin, James Bedford, Ramani Moonesinghe
University College London

Control charts have found increasing use in the monitoring of outcomes from medical treatment. Devices such as exponentially weighted moving average charts (EWMAs) allow clinicians to track their own or their hospital's performance in real time. Statistically defined control limits provide thresholds that help to separate the signal of potential poor performance from the noise of uninformative variation. There are good reasons to think that statistical monitoring can benefit patients: for example, had outcome monitoring been in place, the serial murders committed by English GP Harold Shipman might have been detected earlier. On the other hand, large-scale statistical monitoring systems will inevitably cause some false alarms. If handled insensitively, 'alarms' may induce undue panic and premature attribution of blame that can damage staff morale and engender distrust of the control charts themselves.The statistical literature mainly deals with the technical specifications and distributional properties of control charts. Yet there is little published work on how these charts are used in practice, or how statisticians and non-statisticians ought to collaborate to ensure effective implementation and meaningful interpretation. This talk will present a case study of monitoring patient deaths after emergency bowel surgery in over 150 hospitals in England, which together perform over 20,000 operations per year. We will describe the design of risk-adjusted EWMA charts, discuss challenges of implementation, and present guidelines for clinicians and service managers that propose a course of action to take when a control limit is exceeded. We will also discuss some common misinterpretations of control charts, and the temptation to use them for purposes other than what they are good for. Faced with demands to explain "what the data tell us", a crucial task for the statistician is to communicate uncertainty and to be clear about both the benefits and the limits of what statistics can do.

10.8 RSS Prize Winners: Best presentations from RSC 2019

Thursday 5 September 2pm

Uncertainty Quantification for Offshore Wind Energy

Jack Kennedy
Newcastle University

Energy systems models, are complex, computationally intensive and have a large number of inputs. The model inputs are typically unknowns to be quantified in uncertainty and sensitivity analyses.We represent the unknown quantities by a high dimensional Bayesian prior distribution to be elicited from engineers from academia and industry. From this prior, many samples can be drawn and run through the model to understand the uncertainty induced in the model outputs from the uncertainty in the inputs.However, since such models require vast amounts of CPU time we utilise Gaussian process emulators - statistical approximations to mathematical models - to facilitate computation. Further, we emulate stochastic computer codes with dependent, multivariate output, which adds in an extra layer of complexity in the appropriate description of the dependence and variability in the model outputs.We present the application of emulation techniques to a stochastic offshore windfarm simulation to better the understanding of windfarm output over the early life of the farm; a critical period to ensure the financial viability of offshore wind.

10.8 RSS Prize Winners: Best presentations from RSC 2019

Thursday 5 September 2pm

Improving the Emulation of Stochastic Computer Models

Evan Baker
University of Exeter

Computer models are increasingly being used to learn things about the real world instead of, or as well as, real life experiments. These computer models can take a long time to run, which can make using them for their intended purpose difficult. To help with this, we can create a statistical replacement of the computer model that is much much quicker to run (an emulator), based on a few runs of the computer model. Emulating stochastic computer models can be difficult; requiring a large number of runs from the computer model before the mean and variance can be accurately estimated. I will discuss how deterministic approximations of stochastic computer models can be leveraged to improve the emulation of the stochastic models. Specific implementation details involve the summation of Gaussian processes, which is a popular idea in many areas of emulation.

10.8 RSS Prize Winners: Best presentations from RSC 2019

Thursday 5 September 2pm

Analysis of clickstream data

Ryan Jessop
Clicksco

Online user browsing generates vast quantities of typically unexploited data. Investigating this data and uncovering the valuable information it contains can be of substantial value to online businesses, and statistics plays a key role in this process. The data takes the form of an anonymous digital footprint associated with each unique visitor, resulting in 10^6 unique profiles across 10^7 individual page visits on a daily basis. Exploring, cleaning and transforming data of this scale and high dimensionality (2TB+ of memory) is particularly challenging, and requires cluster computing. We consider the problem of predicting customer purchases (known as conversions), from the customer's journey or clickstream, which is the sequence of pages seen during a single visit to a website. We consider each page as a discrete state with probabilities of transitions between the pages, providing the basis for a simple Markov model. Further, Hidden Markov models (HMMs) are applied to relate the observed clickstream to a sequence of hidden states, uncovering meta-states of user activity. We can also apply conventional logistic regression to model conversions in terms of summaries of the profile's browsing behaviour and incorporate both into a set of tools to solve a wide range of conversion types where we can directly compare the predictive capability of each model. In real-time, predicting profiles that are likely to follow similar behaviour patterns to known conversions, will have a critical impact on targeted advertising. We illustrate these analyses with results from real data collected by an Audience Management Platform (AMP) - Carbon.

11.1 Medical: The latest methodological developments in network meta-analysis

Thursday 5 September 3.30pm

Network meta-analysis of joint longitudinal and time-to-event data

Maria Sudell, Catrin Tudur Smith, Ruwanthi Kolamunnage-Dona
University of Liverpool

Network meta-analysis of joint longitudinal and time-to-event dataJoint models provide methods to analyse potentially linked longitudinal and time-to-event data (termed joint data). Joint data might be longitudinal data complicated by dropout, time-to-event analyses involving time varying covariates, or instances where related longitudinal and time-to-event outcomes are both of interest. An example of joint data is blood pressure repeatedly measured over time, and time until death. Methodology for joint models has been expanded in recent years to meta-analytic joint models, which pool joint data from multiple data sources. The next step in this methodology development is network meta-analytic joint models, which would allow direct and indirect evidence to be utilised when assessing multiple treatment options simultaneously, for related longitudinal and time-to-event data. This talk will describe and discuss aspects of joint models that complicate a joint data network meta-analysis (including time varying components, and relationship between networks of treatment effects for the longitudinal and time-to-event components which are dependent on the joint model association structure), and will introduce modelling approaches to perform network meta-analyses of such joint data.

11.1 Medical: The latest methodological developments in network meta-analysis

Thursday 5 September 3.30pm

Individual Patient Data in a Network Meta Analysis: Is it worth the effort?

Cathal Walsh[2], Joy Leahy[1]
[1]*Trinity College Dublin,* [2] *University of Limerick*

The use of Individual Patient Data (IPD) in Network Meta Analysis (NMA) is becoming increasingly popular. However, as most studies do not report IPD, most NMAs are carried out using aggregate data (AD) for at least some, if not all, of the studies. We investigate the benefits of including varying proportions of IPD studies in an NMA. Several models have previously been developed for including both AD and IPD in the same NMA. We carried out a simulation study based on these models to check the effect of additional IPD studies on the accuracy and precision of the estimates of both the treatment effect and the covariate effect. We also compared the Deviance Information Criterion (DIC) between models to assess model fit. An increased proportion of IPD resulted in more accurate and precise estimates for most models and datasets. However, the coverage probability sometimes decreased when the model was mis-specified. The use of IPD leads to greater differences in DIC, which allows us to choose the correct model more often. We analysed a Hepatitis C network consisting of three IPD observational studies. The ranking of treatments remained the same for all models and datasets. We observed similar results to the simulation study: the use of IPD leads to differences in DIC and more precise estimates for the covariate effect. However, IPD sometimes increased the posterior SD of the treatment effect estimate, which may indicate between study heterogeneity. We recommend that IPD should be used where possible, especially for assessing model fit.

11.1 Medical: The latest methodological developments in network meta-analysis

Thursday 5 September 3.30pm

Reference prediction to connect evidence networks

Howard Thom[1], Joy Leahy[2], Jeroen Jansen[3]
[1] *University of Bristol,* [2] *Trinity College Dublin,* [3] *Precision Health Economics*

In the absence of head-to-head randomised controlled trials (RCTs), network meta-analysis (NMA) can compare treatment effects across a network of connected RCTs. However, RCT evidence for interventions of interest can be disconnected or provide only highly uncertain estimates for relative effects. Single-arm studies, not includable in standard NMA, may also be the only data available on important comparators. In such scenarios, healthcare decision makers, such as the National Institute for Health and Care Excellence in the UK or National Centre for Pharmacoeconomics in Ireland, still need to be able to compare treatment effects.Population adjusted indirect comparison methods have been proposed for comparing disconnected treatments, but these require individual patient data (IPD) which are often held by vested interests. If only aggregate data are available, aggregate level matching (ALM) may be used to connect networks by linking similar RCTs or single-arm studies. A further approach is to predict response of the reference treatment in disconnected RCTs or single-arm studies using a random effect on the baseline treatment in NMA. Random effects on the baseline interfere with randomisation in the connected RCTs and, mostly for this reason, are not generally recommended.We propose a refinement of random effects on baseline, termed reference prediction, to avoid these defects. We build a reference prediction model, using available covariates to improve the predicted response, but keep it separate from the connected RCTs. In the connected network of RCTs, we still assume independent baselines and preserve randomisation. For disconnected RCTs, we re-parameterise treatment effects to be relative to reference and make appropriate multi-arm adjustment for correlations. We demonstrate our method using a real NMA of directly acting oral anticoagulants (DOACs) for atrial fibrillation, which we break into disconnected networks and single-arm studies. We compare reference prediction with both naïve random effects on baseline and ALM.

11.4 Social Statistics: Could I ask you about your life? Future-proofing household surveys

Thursday 5 September 3.30pm

Scottish Government Population Surveys and Core Questions

Ben Cook
Scottish Government

Recently, the four face-to-face population surveys in Scotland were harmonized and pooled together, creating the potential for a combined sample. Statistics from the harmonized subset of questions in the surveys is published as the Scottish Surveys Core Questions. The resulting increase in sample size allows statistics to be reported for minority equality groups (such as specific ethnic and religious groups) and for smaller spatial geographies. Analysis of outcomes for these smaller groups of the population is not possible through individual survey results. The presentation will discuss;- the harmonization of questions,- the coordination of sampling for the individual population surveys, - the development of pooled weighting methodologies for single and multi-year samples,- estimate variability between source surveys. We also present some ongoing results of this unique project, which span multivariate analyses of multi-year datasets, official statistics publications and topic reports. www.gov.scot/sscq

11.4 Social Statistics: Could I ask you about your life? Future-proofing household surveys

Thursday 5 September 3.30pm

Welsh Government National Survey - an example of a merged survey

Steven Marshall
Welsh Government

The Welsh Government and its partner organisations have historically carried out a range of separate surveys of individuals which provide valuable information for policy decisions. Such surveys are expensive to carry out and with downward pressures on cost further exacerbated by falling response rates this prompted a rethink. A review was commissioned in 2014 to consider options that would deliver both robust survey data and achieve value for money. The review was asked to consider all options. As a result five surveys were merged from April 2016 into a single new face-to-face survey – two Welsh Government surveys and three surveys from three separate sponsored bodies. Three of these surveys were conducted face-to-face while one was self-completion and one carried out via telephone. A new combined questionnaire was developed to provide a coherent single whole and designed to try and minimise context effects and discontinuities due to mode changes. The new survey provided significant cost savings of more than £6 million over a 5 year contract period. There have been other benefits beyond cost reduction. Particularly for sponsored bodies the new combined survey allows a more flexible approach to data collection – so that small amounts of data can be collected regularly with more detail every few years rather than a single one off survey every few years. There are improvements in the range of cross topic analysis now possible, from example having sports participation and health topics in the same survey. The basic design of the new survey has retained much of the approach in the previous Welsh Government national survey including the rotation of topics and sub sampling to maximize the full range of topics collected over time as well as being responsive to new data demands.

11.4 Social Statistics: Could I ask you about your life? Future-proofing household surveys

Thursday 5 September 3.30pm

Social Surveys in Northern Ireland - past, present and future

Kevin Sweeney, Andrew McCormick, Brendan Morgan
Northern Ireland Statistics & Research Agency

Household Surveys have come a long way in Northern Ireland (NI), from paper to CAPI, laptop to tablet, floppy disk to 3G, from 1 survey with a 1200 household sample to sampling almost 5% of the addresses in NI annually. Under the last NI Assembly, the development of a 'Programme for Government' modelled closely on the Scottish outcome based approach led to increased pressure for high quality survey data based on probability samples with good response and placing increased demand on survey operations to deliver.This presentation will consider:

- past developments and their implications for the future;

- the current strategy for the development of on-line surveys with a case study of an on-line Census Test;

- some current developments to improve survey efficiency including Achieving Cooperation Training, improving address location, payment of interviewers, use of GPS data;

- the appropriate direction of travel in terms of transformation, translation and offering on-line and mixed-mode options

- is it possible to stem the decline in household survey response

11.5 Methods and Theory: Inferential Machine Learning - Accelerating Statistical Methodology through ML

Thursday 5 September 3.30pm

RKHS-based tests for Survival Analysis

Tamara Fernandez
University College London

Hypothesis testing is a fundamental problem in Machine Learning and Statistics. Over the past decade, statistical tests based on embeddings over reproducing kernel Hilbert spaces (RKHS) have become popular due to their excellent results when dealing with different types of complex data. With the incorporation of new sampling methods in clinical research, it is important that Survival Analysis methodologies are able to deal with complex data, thus the application of RKHS-based methods becomes timely and relevant. While the RKHS methodology has been successfully applied to different problems in statistics and Machine learning, up to the best of our knowledge, it has not been systematically explored in the context of censored data. In this talk we explore different approaches to extend RKHS-based tests to censored data, focusing particularly in the general problem of testing for independence between right-censored survival times and high dimensional covariates. As opposed to the classical/uncensored approach of embedding probability distributions onto an RKHS, we propose to embed hazard functions which are estimated from the data using the Nelson Aalen estimator. Following this approach, we derive an analytically tractable test-statistic and a computationally efficient Wild Bootstrap procedure to estimate the null distribution quantiles. We perform an extensive empirical evaluation of our test-statistic for proportional and time-dependent hazard functions showing excellent results in both settings. We finalise by discussing some theoretical properties and by establishing connections to classical approaches such as log-rank tests and the Cox-Score test.

11.5 Methods and Theory: Inferential Machine Learning - Accelerating Statistical Methodology through ML

Thursday 5 September 3.30pm

Using variational autoencoders to learn efficiently embedded representations of functions and their properties

Swapnil Mishra, Seth Flaxman, Samir Bhatt
Imperial College London

Many machine learning methods specify a function class and an algorithm to learn the "best" function in that class, for example by minimizing the empirical risk. Function classes can range from simple linear models to reproducing kernel Hilbert spaces (RKHS). For a number of popular algorithms, such as Gaussian process regression, learning from the function class comes with computational challenges, including computational burden, correlated parameters and algorithmic complexity. Here we use variational autoencoders (VAE) to learn low dimensional embeddings of function classes. We show that our VAE framework can accurately learn complex function classes such as Gaussian, log-Gaussian Cox, and Hawkes processes. We also show that our VAE framework can be used efficiently in both Bayesian and non-Bayesian inference schemes.

11.5 Methods and Theory: Inferential Machine Learning - Accelerating Statistical Methodology through ML

Thursday 5 September 3.30pm

A unified machine learning approach to time series forecasting applied to emergency department demand

Michaela Vollmer, Samir Bhatt, Seth Flaxman, Graham Cooke, Ben Glampson, Luca Mercuri, Swapnil Mishra
Imperial College London,

There were 24.8 million attendances at Accident and Emergency (A&E) departments in England in 2018-2019 corresponding to an increase of 11 million attendances in the past ten years. This steadily rising demand at A&E departments leaves the NHS under constant pressure to provide a good quality of care while being highly productive. Managing hospital demand efficiently requires an adequate knowledge of the future rate of admission. Using admissions data for the past eight years from St Mary's and Charing Cross Hospitals in London, we developed a novel predictive framework to understand the temporal dynamics of hospital demand and we applied an exhaustive statistical analysis to daily presentations at the A&E departments at St Mary's and Charing Cross Hospitals, evaluating a range of standard time series and machine learning approaches. Ultimately, we applied a novel ensemble methodology to combine the outcomes of the best performing time series and machine learning approaches in order to make very accurate forecasts of demand, 1, 3 and 7 days in the future. We found that while St Mary's and Charing Cross Hospitals both face an average daily demand of 208 and 106 respectively and despite considerable volatility around this mean, we were able to predict attendances at these emergency departments up to a mean absolute error rate of 7 patients corresponding to a mean absolute percentage error of around 3% which may help to save up to £900,000 at the A&E department at St Mary's Hospital alone every year.

11.6 Communicating & Teaching Statistics: Do we need a Q-Step initiative for statistics training in biology, medicine and health?

Thursday 5 September 3.30pm

Adapting the Q-step scheme for use in undergraduate Medicine: What would we need to do?

Margaret MacDougall
University of Edinburgh

The success of the Q-step scheme in improving employability for social science graduates intent on pursuing careers requiring quantitative research skills has been recognized. This resource-intensive venture has been made possible through generous funding from the Economic and Social Research Council. The cross-institutional nature of the work provides scope for exploring a sustainable model where good practice and content from educational programmes are shared across institutions. The relevance of quantitative learning is reinforced through provision of Q-step work placement schemes, which invite students, as employees, to experience authenticity when applying quantitative, including statistical, skills to solve genuine real-life problems. This paper briefly explores the 'why?' and 'how?' of developing a similar initiative in statistical learning, with a focus on undergraduate medical education. This will provide an opportunity to discuss the key players in approving requirements for entry to medical school and determining required learning outcomes for UK medical schools to deliver in ensuring the clinical competence of their new graduates. In supporting the case for a medical statistics variant of the Q-step scheme, I will suggest recently reported funded research and recommend, and open the floor for further discussion on, potential funding sources to explore. However, I will also bring to the table an honest impression of some key challenges which need to be tackled collectively across prospective participating institutions. Further, I will acknowledge the opportunities and challenges presented by a crowded medical curriculum, particularly in relation to guaranteeing placement opportunities with adequate statistical content for all. Additionally, I will highlight a range of resource challenges in attempting to adopt an integrative approach to statistical learning throughout the medical curriculum, not to mention the need to convince curriculum managers that the efforts are worth investing in. Delegates will be welcome to share their viewpoints in response to the above proposal.

11.6 Communicating & Teaching Statistics: Do we need a Q-Step initiative for statistics training in biology, medicine and health?

Thursday 5 September 3.30pm

How could we design a Q-Step initiative for biology, medicine and health?

Jamie Sergeant
University of Manchester

If we accept the premise of the session's title, that we do need a Q-Step initiative for statistics training in biology, medicine and health, or at least that such an initiative would be desirable, how might we go about designing one? This part of the session will explore ideas and preferences from delegates on what the objectives of a Q-Step-type initiative in this field might be and how the statistics community could work together to achieve them.

Poster viewing A

Wednesday 4 September 1.30pm

Bayesian Networks to assess the impact of 20mph speed limit policies on road traffic collisions and casualties

Glenna Nightingale[1], Ruth Hunter[2], Frank Kee[2]
[1] *University of Edinburgh,* [2] *Queen's University Belfast*

This study involves the use of Belief Networks for the evaluation of the impact of the 20mph speed limit policy in the cities of Edinburgh and Belfast. Belief Networks are independence probability graphs which consist of nodes (key variables) and arcs (hypothesized links between key variables). We propose the use of multiple logic models in combination with the results from a "contribution trial" to construct these Belief Networks. Essentially, a contribution trial is a workshop which brings together key domain experts to seek their views on the causal pathways inherent in the project's logic models. The Belief Networks will be updated using empirical "evidence" from our speed/casualty/collisions data streams to generate posterior probabilities of key outcomes. We aim to perform model discrimination in a Bayesian framework to determine which Belief Network or cluster of Belief Network receives the highest probabilistic support. The Belief Network with the highest probabilistic support, and highest practical significance will be used for inference. We will consider gateway strategies (Dmitrienko et.al, 2011) where relevant as much as possible. The Belief Network with the highest probabilistic support and practical significance will be incorporated into a Web app (using R Shiny) to make the compiled network accessible to policy makers and public health professionals. Empirical evidence such as change in average speed, or increased speed limit enforcement can be entered into the app, and the change in the conditional probabilities of key outcomes such as traffic collision rates and casualties are outputted.Overall we aim to develop a robust statistical tool using domain expert information (in a contribution trial) and Bayes probabilities to equip decision makers in assessing the impact of the complex intervention of the 20mph speed limit policy. The method developed will be encapsulated in a Web app making the method easily accessible. The initial prototype of the proposed app is now available for discussion.

Poster viewing A

Wednesday 4 September 1.30pm

GEOTECHNICAL PROPERTIES OF SOIL IN SOLID WASTE DUMP SITE

Micheal Olobadola
Funaab

Soil samples at various depths from different sampling points on an open dumpsite were assessed in order to investigate the impact of municipal solid waste dumping activities on some geotechnical properties of soil. The geotechnical parameters, which include moisture content, permeability, grain size analysis, compaction and Atterberg limits, were assessed using ASTM standard. Results showed that geotechnical parameters for dumpsite soils are higher than the control samples. The liquid limit (LL) and plasticity index (PI) ranged from 26-38 and 11-20 respectively for dumpsite soil while the mean values of LL and PI for control soil samples were 31 and 16 respectively. Soil moisture content was significantly higher in dumpsite soil (25.8 – 33.6%) than control soil (10.7%). The percentage passing (N200) sieve test ranged from 2.9 – 19.5% for dumpsite soil and 4.9 - 18.8% for control samples. The soil under the control treatment had least permeability value of 0.0043 cm/sec while a maximum value of 0.0058 cm/sec was obtained in the dumpsite. Maximum Dry Density (MDD) and Optimum Moisture Content (OMC) for dumpsite ranged from 1.5 – 7.7g/cm3 and 1.4 – 29.11% respectively. However, the MDD and OMC values for control soil were 1.3g/cm3 and 6.6% respectively. The settlement (compressibility) analysis revealed that soil samples in the study area are of low/medium degree of expansion and non critical/marginal degree of severity status for stability purpose. The Compression Index (CC) values ranged from 0.12 – 0.25, suggesting low to moderate compressibility over an engineering time. Most of analyzed geotechnical parameters values fall with the specification limits for construction purpose. The result obtained also showed that dumping activities have great effect on soil's geotechnical properties. The obtained geotechnical parameters results for soil samples in the study area showed

Poster viewing A

Wednesday 4 September 1.30pm

Data Fusion of Cardio-Metabolic Risk Prediction for Cognitive Impairment in the Middle-Aged: the CARDIA Study

Sujin Kang
Imperial College London

Background: There is a lack of evidence of the role of cardio-metabolic risk factors in brain health from large cohort studies which are comprised of metabolites; specifically, from a life-course perspective with increasing CVD risk profiles.

Aim: The primary aim of this study is to investigate associations between cardio-metabolic risk factors and cognitive impairment with brain MRI-derived parameters in the Coronary Artery Risk Development in Young Adults (CARDIA) Study at Year 30, and to investigate if targeted urinary metabolites are mediators of the associations. Significant epidemiological exposures and targeted urinary metabolites will be examined, and the information will be used to develop a risk assessment tool for brain health (i.e., cognition).

Design and population: This present study focuses on the CARDIA Study sub-cohort of 606 participants in middle-aged (aged 48-60 years who completed cognitive tests) with high-resolution nuclear magnetic resonance (NMR) spectroscopy data, 340 of them with liquid-chromatography mass spectrometry (LC-MS) data and 280 of them completed the brain MRI data at Year 30, 2015-16.

Methods: Bayesian models and advanced network classification models will be employed to investigate metabolic phenotypes. Different types of significant (1) clinical, socio-demographic, lifestyle risk factors and (2) metabolite markers which will be integrated (i.e., fusion) to provide a cardio-metabolic risk-specific integrated map of brain health. Longitudinal analysis of the associations will be used to explore the possibility of developing a novel, comprehensive risk assessment tool of brain health.

Discussion: This study will employ a novel biostatistical approach of a fusion analytic method for the different types or levels data on the large-scale epidemiological study. The approach may be able to recognise complex multivariate epidemiologic, pathologic and phenotypic relationships across the disease network that are yet unidentified.

Poster viewing A

Wednesday 4 September 1.30pm

Statistical Analysis on Industrial Sector of Nigeria Economy

Iyabode Favour Oyenuga
The Polytechnics, Ibadan, Nigeria

Industrial sector generates revenue, creates services, brings about incomes and create employment. Over the years, the Nigerian economy has recorded a decline in industrial growth with some industries recording closures as a result of difficult operating environment. Few studies have examined the role of industry in economic growth of Nigeria. The study evaluated the industrial sector of Nigeria gross domestic product with the use of Secondary data over a period of 27 years from 1991 to 2017 extracted from Central bank of Nigeria bulletin.Multiple regression method was adopted to analyze the data in order to test for individual and joint parameter, heteroscedasticity, multicollinearity, autocorrelation and predictive power. The results revealed that all the explanatory variables contribute significantly to the model. There is a very strong positive correlation between the dependent variable (Industry) and each of the explanatory variables (Crude Petroleum and Natural Gas, Solid Minerals and Manufacturing). The model has high predictive power of 94.4%, that is, 94.4% of the total variation of the dependent variable is explained by the regression model.It was recommended that government should encourage the growth of industries in Nigeria through establishment of favourable policies for industrial development.

Poster viewing A

Wednesday 4 September 1.30pm

Modelling and monitoring the environmental impacts of aquaculture

Michael Currie[1], Marian Scott[1], Claire Miller[1], Andrew Berkeley[2]
[1] *University of Glasgow,* [2] *Scottish Environment Protection Agency*

Aquaculture accounts for nearly half of the global fish supply and is an expanding industry around the world. Current farming techniques deposit a range of wastes on the seabed which are dispersed, creating a zone of impact that is monitored to ensure that agreed quality standards are met. In Scotland, the Scottish Environment Protection Agency (SEPA) use a particle tracking model called NewDEPOMOD to predict the impact of aquaculture on the seabed. When investigating the performance of NewDEPOMOD, two challenges are presented: (i) limited observational data are available to validate the NewDEPOMOD predictions and (ii) the computational cost of running NewDEPOMOD.There is some uncertainty as to the suitability of some of the default parameter values within NewDEPOMOD, and hence a sensitivity analysis has been developed to assess the uncertainty of these parameters and the impact on the NewDEPOMOD predictions. this work will then be used to inform the development of a statistical emulator to assess the environmental impacts of aquaculture without the computational cost of NewDEPOMOD.The sensitivity analysis was conducted at 2 sites with different physical properties. In collaboration with SEPA, 11 parameters were identified for the sensitivity analysis, and plausible ranges for their values were determined based on literature and expert knowledge. Latin Hypercube Sampling was used to create 100 different sets of values for the parameters, accounting for interactions that exist between some of the chosen parameters, and 100 runs of NewDEPOMOD were performed to account for a random walk element within NewDEPOMOD. A Random Forest approach was then used to investigate which parameters described most of the variation in the NewDEPOMOD predictions.It is clear from the sensitivity analysis that the physical properties of a site play a key role in the influence of the parameters on the NewDEPOMOD predictions.

Poster viewing A

Wednesday 4 September 1.30pm

Response Surface Approach to Optimizing Baking Parameters for Improved Quality Loaf with Local Content in Nigeria

Polycarp Chigbu
University of Nigeria Nsukka, Nigeria

Objective: A study was proposed to determine the optimum combination of imported wheat and local cassava flours, baking time and temperature. The aim is to increase local content in bread production in Nigeria while attaining acceptable quality of bread loaf. The three variables under consideration are Baking time, (in minutes), Baking temperature, (in degree Celsius) and Wheat/Cassava mix concentration, (in percentage). Three important qualities of baked loaf that were to be monitored are moisture content, (in percentage), volume of loaf, (in cm3) and plastic limit, (in cm).

Method: Response surface methodology was considered appropriate since the study requires the optimization of baking factors and the central composite design was used for that purpose. Different variations of the CCD involving partial replications were applied and evaluated using the fraction of design space graphs to identify, among the numerous design options, the design that gives the best spread of minimum variance of prediction of responses throughout the design region. Also, five axial distances, , were considered: the spherical cuboidal , practical , rotatable and orthogonal .

Results/Conclusion: The CCD with cuboidal axial distance and with the star portion replicated twice is recommended for the experiment.

Poster viewing A

Wednesday 4 September 1.30pm

Detection of voids in additive manufacturing using x-rays in projection space

Sherman Lo[1], Julia Brettschneider[1], Thomas Nichols[2], Jason Warnett[3], Gregory Gibbons[3], Mark Williams[3]
[1] *University of Warwick,* [2] *Oxford Big Data Institute,* [3] *Warwick Manufacturing Group*

X-ray computed tomography (CT) can be used for defect detection in additive manufacturing. The object is scanned at multiple angles to reconstruct the object in 3D space. The process can be time consuming. The aim is to investigate if it is possible to conduct defect detection from a single scan to speed up the quality control procedure. An experiment was conducted, a 3D printed sample was manufactured with voids to see if they can be detected.Photons behave randomly. Hence to do defect detection pixel by pixel, uncertainty must be taken into account. A compound Poisson model was used to model the grey values in a pixel. It assumes that photon arrivals are a Poisson process with Gamma distributed energy. This resulted in a linear relationship between the mean and variance of the grey value, which can be used for variance prediction and to quantify the uncertainty.Software was used to simulate the scan and it was compared with the x-ray acquisition under the face of uncertainty. Each pixel was treated as a hypothesis test. The software, and the information provided to it, was not perfect which led to model misspecification and incorrect inference. The empirical null filter was proposed. The filter locally normalise each test statistic using robust estimators of the mean and deviation. This reduced the number of false positives. Sensible inference was achieved on the dataset provided.

Poster viewing A

Wednesday 4 September 1.30pm

Geometric Characterizations and Symmetric Relations between Standard Normal Distribution and Inverse Mills Ratio based on Pythagorean Theorem

Shingo Nakanishi[1], Masamitsu Ohnishi[2]
[1] *Osaka Institute of Technology / Computing Center,* [2] *Osaka University / Graduate School of Economics and Center for Mathematical Modeling and Data Science*

We dealt with the geometric characterizations between standard normal distribution and inverse Mills ratio by circle and square from the viewpoints considering the height of densities such as the ancient Egyptian drawing pictures by using the Greek Pythagorean Theorem. First, we can clarify that the general solution as the second order differential equation of the integral form of cumulative distribution function of standard normal distribution and that of Bernoulli differential equation of inverse Mills ratio are tied on the rotational symmetric figure. Especially, we can also confirm the integrals of cumulative standard normal distribution are related to inverse Mills ratio geometrically. Second, from these tendencies, we can also get the geometric and symmetric modified intercept forms for maximal profits of winners, these losses of losers, and their banker's fee. We can understand that these equations are changing such as Pythagorean Theorem correctly based on the relations between the probability points and these probabilities. If the probability is 0.612003 which was found by Karl Pearson, we can show you that it is the special point of standard normal distribution such as Sir David Roxbee Cox's proposal. The right triangle with 3:4:5 is also the third quantile of standard normal distribution. Third, we can also realize there are many similar tendencies close to the relations between circles and squares such as Vitruvian man by Da Vinci and Mandalas although there might not be related to normal distribution directly and historically. The ancient Egyptian drawing styles enable us to illustrate symmetric relations and geometric characterizations between standard normal distribution and inverse Mills ratio with circle and square based on Pythagorean theorem. We think that our ideas shall be contributed in the statistical modelling and these evaluation fields since our suggested figures might be more easily to be understood and powerful than we thought.

Poster viewing A

Wednesday 4 September 1.30pm

When impairment is a handicap: higher education in Brazil

Kaizo Beltrao[1], Hugo Simas[2], Moema Teixeira[2], Ricardo Goes[2]
[1] *EBAPE FGV,* [2] *Cesgranrio Foundation*

In 2016, the Brazilian Government passed a Law (13409) defining quotas for ethnic groups and individuals with disability. This Law was actually a revision of a previous legislation, Law 12721, enacted in 2012. Since 1995, Inep (a department of the Ministry of Education) has been collecting data on University courses and students enrollment through the Brazilian higher education census. From 2009 on, this census individually lists and identifies all students enrolled in University and College courses with information on age, gender, disability, enrollment status, race/ethnic group, majors, scholarship and selection method for entry in the course (e.g. through a national government exam, private exam, quotas). With the student identification number, it is possible to follow students across different years and check if they are still enrolled, if they droped-out, if they got transferred to another course, if they died, if they are licensed, if they graduated, among other possibilities.This text intends to follow up students with different disabilities. The anual higher education census lists thirteen different disabilities: blindness, partial visual impairment, Deafness, partial hearing impairment, physical deficiency, deafness combined with blindness, multiple deficiency, intellectual deficiency, Asperger syndrome, Autism spectrum disorder, Rett syndrome, and so on. Drop-out rates and average time to completion for each disability group and the overall population controlling for availability of in campus resources for those with special needs. We find out that students with special needs do take, in average, longer to graduate than other students, but the expected time is mediated by the existence of in-campus resources and programs. Some regions in Brazil are better equipped with these resources. Data is modelled using a survival model.

Poster viewing A

Wednesday 4 September 1.30pm

New Second-Order Asymptotic Methods for Nonlinear Models

Gubhinder Kundhi[1], Paul Rilstone[2]
[1] *Memorial University of Newfoundland,* [2] *York University*

New higher-order asymptotic methods for nonlinear models are developed. These include generic methods for deriving stochastic expansions of arbitrary order, methods for evaluating the moments of polynomials of sample averages, a method for deriving the approximate moments of the stochastic expansions with simplified expressions for the first four moments and a third-order approximate Saddlepoint expansion. These techniques are applied to improve inferences with the Two-Stage Least Squares estimator and the weak instruments problem. It is well established that Instrumental Variable (IV) estimators in the presence of weak instruments can be poorly behaved, in particular, be quite biased in finite samples. In our application, finite sample approximations to the distributions of the IV estimators are obtained using Edgeworth and Saddlepoint expansions. Higher order analytical corrections provided by these expansions are used to analyze departures from normality. In a Monte-Carlo experiment, the performance of these expansions is compared to the first order approximation and other techniques commonly used in finite samples such as the bootstrap.

Poster viewing A

Wednesday 4 September 1.30pm

ON THE USE OF VARIABLE TRANSFORMATION IN ESTIMATING POPULATION PARAMETERS OF THE STUDY VARIABLE

Chinyeaka Izunobi, Aloysius Chijioke Onyeka
FUTO, Federal University of Technology, Owerri. Imo State, Nigeria

This work is based on the use of variable transformation in estimating population parameters of the study variable in Stratified random sampling scheme. The properties of the estimators were obtained. Numerical illustration was carried out to verify theoretical results in the literature and to compare the efficiencies of some selected estimators that utilized variable transformation. Results of the study revealed that the use of variable transformation often leads to increased efficiency over the sample mean estimator. It was also observed that when using variable transformation, increased efficiency could be achieved by increasing the number of transformed variables to be used in a given survey.

Poster viewing A

Wednesday 4 September 1.30pm

Modelling Baseline Hazards in a Cox Model using Gaussian Processes

Bindu Vekaria, Glen Martin, Thomas House
University of Manchester

There is growing interest in developing prognostic models that produce absolute measures of risk for an outcome of interest (e.g. time-to-death). These provide an effective way of communicating risks with patients and aid decision-making. In the context of time-to-event outcomes, such models are usually based on the Cox model, where parametric distributions (e.g. Weibull) are used to model the baseline hazard. However, these lack the flexibility needed to model complex time-varying hazards. Alternatively, flexible parametric methods model the baseline hazard using cubic splines, but such models are difficult to validate, rarely used in practice and difficult to interpret.Joensuu et al. (2012) proposed an alternative method utilising a Bayesian framework by modelling the log baseline hazard with a Gaussian Process. While this is attractive since it models the baseline flexibly, it has only been applied to scenarios where one can assume a piecewise constant hazard. Generally, this is not representative of clinical settings.Therefore, we propose a novel piecewise linear Gaussian Process, which can be extended to higher order polynomials, where necessary. Here, I present the results from a simulation study that compares the predictive accuracy of our method with existing methods, across a range of representative scenarios.

Poster viewing A

Wednesday 4 September 1.30pm

R: A Hitchhikers Guide to Reproducible Research

Brendan Palmer, Darren Dahly, Joseph Eustace
HRB Clinical Research Facility Cork

As funding agencies increasingly require adoption of open science practises in the research that they fund, there is a pressing onus on researchers to respond. For many, awareness of the tools, skillsets and resources available to achieve this is lacking. University subscriptions to statistical analysis software, such as (but not limited to) SPSS, restricts the end user to a defined suite of operations and functions, that may not be accessible at a later stage in the research project or the researchers career. Furthermore, such programmes do not allow for reproducible research to be faithfully applied.Conversely, the R programming language is free to download, install and use. Together with the graphic user interface, RStudio, it is now the favoured data analysis tool across a variety of fields, including statistics. Advances in R software development, namely the reorganisation of popular R packages under the umbrella banner of the “tidyverse”, has made R accessible to a wider audience. This has numerous benefits, chief amongst them being that R, as a programming language, has become more human readable, intuitive and accessible. Data importation, cleaning, exploration, visualisation, analysis and reporting can be enacted through R. Combined, R allows for reproducible workflows to be enacted, packaged and published.We conclude that senior academics and programme coordinators adopt R modules as a core element of academic programmes aligned with research activities. We propose this be implemented through structured PhD training modules or as part of early career researcher continuing professional development.

Poster viewing A

Wednesday 4 September 1.30pm

Variance Estimation in Dual Channel Microarray Applied on Split-Plot Design

Abimibola Oladugba, Ifeoma Ogochukwu Ude
Department of Statistics, University of Nigeria, Nsukka

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, color, first and second levels of treatment B at each level of treatment A and first and second levels of treatment A at each level of treatment B 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 for comparing the effects of treatment comparison.

Poster viewing A

Wednesday 4 September 1.30pm

Major Errors in Research on Lighting and Public Safety

Paul Marchant
Leeds Beckett University

The poster gives a couple of examples of where statistical errors give rise to claims of detecting very large beneficial effects in an area of public policy. However, the serious errors committed coupled with the absence of checkable data, make such positive claims suspect. The poster points out where the major errors reside and indicates, through detective work, the consequences for the claims.

Poster viewing A

Wednesday 4 September 1.30pm

Interpretable drivers of sensitivity analysis for non-ignorable missing covariate in linear regression models

Peng Yin[1], Rong Zhu[2], Jianqing Shi[2]
[1] *Chinese Academy of Sciences,* [2] *University of Newcastle upon Tyne*

Goodness-of-test is not valid for non-ignorable missing data problems because of non-identiability. The model structure for missing data is very complex and in many cases, both ignorable and non-ignorable missingness mechanisms are plausible. Sensitivity of missing data mechanism demands serious circumspection. In this paper we will utilize a likelihood-based analysis platform to investigate the sensitivity of missing data mechanism (MDM) modelling in the missing covariate problems. Our analysis will adopt an approximate measure of incomplete data bias for parameter estimations, which characterizes the magnitude of uncertainty and its gradient from a specied model. We will discuss the sensitivity measures in linear regression models with missing covariate specically. How to identify key interpretable sensitivity parameters is very important and this will be addressed in quite details. Our discussion will show a well approximated and easy interpretable formular for measuring incomplete data bias. Numerical examples will be presented in simulation studies and real data example.

Poster viewing A

Wednesday 4 September 1.30pm

A randomized controlled trial of a school-based smokeless tobacco intervention to improve knowledge and prevent adolescent tobacco use

Shafquat Rozi[1], Nida Zahid[1], Talat Roome[2]
[1] *Department Of Community Health Sciences, Aga Khan University, Karachi, Pakistan,* [2] *Dow University of Health Sciences, , , , , ,*

Objectives: To assess the effectiveness of intervention in improving knowledge, attitude and perception regarding smokeless tobacco (SLT) use and its harmful effects and intention to quit SLT among school going adolescents.

Methods: A school-based cluster randomized control trial was carried out in 18 secondary schools targeting male and female students from grades 6 to 10 in Karachi. Intervention comprised of a 30 minutes power point presentation, two posters, and one pictorial booklet and a video game on the hazards of use of various tobacco products. Primary outcome was knowledge about hazards of smokeless tobacco (SLT) and secondary outcome measures were 1) Attitude about hazards of SLT and 2) Perception about hazards of SLT.

Results: We enrolled 738 participants in intervention group and 589 in the control group. Mean score of knowledge significantly improved in intervention as compared to control group (p value <0.01). Intention to quit was found to be proportionately higher (33%) in the intervention group as compared to control group. Generalized estimating equations (GEE) were used to assess the association of various factors with knowledge regarding harmful effects of SLT use. Significant predictors of increase in knowledge score were found in children: who had seen any anti SLT messages on social media in the past 30 days, who were getting information regarding harmful effects of SLT use in school or textbooks and who had friends using SLT.

Conclusions: A school-based intervention was effective in increasing knowledge regarding the harmful effects of SLT use and intention to quit smokeless tobacco use among school going adolescents. Introduction of such educational programmes on a regular basis in schools or as part of school curriculum can have an impact on reducing prevalence of SLT use.

Key words: Smokeless tobacco, cluster randomized trial, adolescents, school based intervention

Poster viewing A

Wednesday 4 September 1.30pm

Multilevel weighted ordinal regression analysis to determine the association of bullying with depression symptoms among school adolescents of Pakistan

Shafquat Rozi, Maryam Ali, Apsara Ali, Wajeeha Zahid
Department Of Community Health Sciences, Aga Khan University, Karachi, Pakistan

Purpose: In poor resourced country, youth violence is the most neglected problem to be focused because of its low reporting rate and less awareness of it effect on psychological wellbeing of adolescents. The aim of this study is to determine the association of bullying status and other factors with depression symptoms among school going adolescents of Pakistan.

Methods: We used secondary data from The Global School-based Student Health Survey (GSHS) conducted in 2009. This was a school-based cross-sectional survey conducted on nationally representative sample of adolescents from 8th to 10th grades with age 11–16 years. A two-stage cluster sampling technique was employed for selection of school and children. The outcome variable was presence of depressive symptoms whereas, the primary exposure was bullying victimization experienced by adolescents in last 30 days. Multilevel weighted ordinal regression analysis was performed.

Results: A significant proportion of severely (48%) and moderately (46%) depressed adolescents reported being bullied. Whereas, three fourth of mildly/non-depressed adolescents (70%) reported having no exposure to bullying in past 30 days. A significant interaction was found between bully and parent-child relationship indicating that adolescents exposed to long term bully and have poor parent-child relationship had 3 times higher odds of depression (OR: 3.35, 95% CI: 1.95, 5.75) as compared to adolescents with no exposure to bully with a good parent-child relationship.

Conclusion: This particular study has identified the potential risk factors that will help in formulating preventive strategies to address the toll of depressive symptoms experienced by adolescents.

Poster viewing A

Wednesday 4 September 1.30pm

Complex interaction analysis by liquid association and its extension

Kerchau Li
Institute of Statistical Science, Academia Sinica

Data science has been fostered by the rapid accumulation of a growing number of large open databases, each containing tens of thousands of variables pertinent to the knowledge advance of its specific scientific domain. For example, in cancer research, the Cancer Genome Atlas(TCGA) consortium has completed its mission of molecular-profiling over 11,000 tumors from 33 of the most prevalent forms of cancer. Here a general statistical framework to enhance the information-distilling from the complex patterns of interaction between variables is presented. We have earlier introduced Liquid Association (LA) for inferring higher order of association between variables in a system (Li 2002, Proceedings of National Academy of Sciences, USA). LA was originally introduced to study patterns of gene-gene interaction that involve three genes at a time. LA aims at finding how the correlation pattern between a pair of functionally associated genes may be dynamically altered due to the influence of ever-changing but often-hidden cellular state. It is computationally expensive to compute LA scores for all possible triplets in very large datasets, but Graphic Processing Units (GPUs) can be used to speed up the LA computing. There are several ways of extending LA, including background adjustment and dimension reduction. We shall demonstrate how they can be used to exploit complex interaction and broaden the scope of statistical multivariate analysis.

Poster viewing A

Wednesday 4 September 1.30pm

Likelihood approximation and prediction for large spatial and spatio-temporal datasets using hierarchical matrices

Anastasiia Gorshechnikova[1], Carlo Gaetan[2]
[1] *University of Padova,* [2] *Ca' Foscari University of Venice*

Large datasets with irregularly sited locations are difficult to handle for several applications of Gaussian random fields such as maximum likelihood estimation (MLE) and kriging prediction due to a high computational complexity. For relatively large spatial and (or) temporal dimensions the exact computation becomes unfeasible and alternative methods are necessary. Several approaches have been proposed to tackle this problem. Most of them assume a specific form for the covariance function and use different methods to approximate the resulting covariance matrix. The aim is to approximate the covariance functions in a format that facilitates the computation of MLE and kriging prediction with large spatial and spatio-temporal datasets. For a sufficiently general class of spatial covariance functions, a methodology is developed using hierarchical matrices. This technique involves the partitioning of a matrix into sub-blocks according to specific given conditions with the further approximation of the majority of these sub-blocks by low-rank matrices. The resulting covariance matrix allows for computation of the matrix-vector products and matrix factorisations in a log-linear computational cost followed by an efficient MLE and kriging prediction. Numerical experiments are performed on simulated spatial and spatio-temporal data to recover the true values of the parameters of the covariance matrix. This method is further applied on a real dataset consisting of measurements of atmospheric carbon dioxide mole fractions. The prediction accuracy and computational time are compared with other methods. As a result, among the methods considered in this study, the approach presented is the most efficient in terms of the root mean-squared prediction error and computational time.

Poster viewing A

Wednesday 4 September 1.30pm

Predictive Modeling for Nursing-care Levels and Costs

Yasutaka Hasegawa[1], Hideyuki Ban[1], Tetsuya Moriike[2], Hayato Nishikawa[2], Satoshi Horie[2]
[1] Hitachi,Ltd. Research & Development Group, [2] Hitachi Ltd. Government & Public Corporation Information Systems Division

Background: The growth of nursing-care cost, which is caused by aging population, is one of the most critical issues in Japan. Each local government is required to plan effective nursing-care prevention measures utilizing healthcare data, and it is important to plan prevention measures that consider not only the current nursing-care situation but also the future.

Objectives: In this study, we created a model to predict future nursing-care levels and costs using nursing-care, medical and checkup data.

Method: First, using machine learning, we created a predictive model for future nursing-care levels from 487 explanatory variables included in the data. Next, using a linear regression, we created a model that predicts future nursing-care costs from the prediction result of nursing-care levels and current nursing-care costs. To create a model for future nursing-care levels, the following seven machine learning methods were tried; (a) Binomial logistic regression, (b) Multinomial logistic regression, (c) Ridge regression, (d) Lasso regression, (e) Elastic Net, (f) Gradient boosting, and (g) Support vector machine.

Results: We used data of 405,687 people from 2014-2015 to create predictive models. As a result of the evaluation using test data of 81,136 people, it was confirmed that model (a), (d), (e), and (f) can predict nursing-care levels after one year with more than 95% accuracy and that nursing-care costs after one year can be predicted with less than 1% error. In addition, the model (a) can analyze explanatory variables contributing to the occurrence and progress of nursing-care, and it found that checkup and nursing-care services reduce the occurrence and progress of nursing-care.

Conclusions: The model (a) is a high accuracy model that can explain factors that contribute to the occurrence and progress of nursing-care, and confirmed that this model is the most appropriate model for planning nursing-care prevention measures in the local government.

Poster viewing A

Wednesday 4 September 1.30pm

HyCOSM: a Hybrid-COupled-Stochastic Model for subseasonal climate prediction

Michel d. S. Mesquita[12], Michel d. S. Mesquita, George U. Pedra[3], Saurabh Bhardwaj[4], Santosh K. Muriki[4], Lincoln M. Alves[3], Iracema F. A. Cavalcanti[3]
[1] Bjerknes Centre for Climate Research, [2] Future Solutions, Norway, [3] National Institute for Space Research, INPE, Brazil, [4] The Energy and Resources Institute, TERI, India

The Intraseasonal scale is particularly important in climate science because it acts as a modulator of weather systems. It is also one of the most difficult time scales to predict due to the strong influence of large-scale atmospheric systems. We have developed a subseasonal forecast model called HyCOSM: a Hybrid-COupled-Stochastic Model for subseasonal prediction. HyCOSM allows for the prediction of extreme precipitation of up to 28 days. It consists of three distinct steps: the first (S1) is the characterization of the intraseasonal climatic pattern and the development of a diagnostic index through Maximum Covariance Analysis (MCA); we call this step the Multivariate Intraseasonal Rainfall Index (MIRI). This index is made of two scalar components, which are then modelled in the second phase (S2) using a Vector Autoregressive (VAR) model. For the third step (S3), we have used Artificial Neural Networks (ANN) to obtain the intraseasonal rainfall forecasts. Results from S1 show that the MCA output captures 71% of the intraseasonal rainfall variability in India and 65% over South America; it also represents reasonably well the rainfall signal, as well as the atmospheric circulation and convective patterns; Furthermore, it describes the evolution mechanism of atmospheric systems in both space and time, making it possible to create a projection in a space-phase diagram. In S2, the VAR model captures the variability of each scalar component up to 28 days in advance. The correlation coefficient between the forecast and the observed index is larger than 0.7 ($p < 2.2e-16$). In S3, the ANNs provide a 28-day rainfall forecast of extreme precipitation. Overall, the information generated through HyCOSM can significantly help decision makers and farmers regarding extreme events, short-term precipitation trends, and the evaluation of other rain-related variables.

Poster viewing A

Wednesday 4 September 1.30pm

Multi-step Regression Model for Estimating Hospital Length of Stay

Hirofumi Kondo, Kaoru Kobayashi, Yasutaka Hasegawa, Hideyuki Ban
Hitachi,Ltd. Research & Development Group

[Background and Objectives] The growth of medical cost caused by aging and increasing of chronic disease patients has become one of the critical issues in the developed nations. To solve this isuue, measures focusing on health promotion and disease prevention are important. This time, in order to visualize disease risk and utilize it for the health business, we developed a model that predicts the number of hospitalization days for the 8 major lifestyle-related diseases based on medical check-up results and past medical history.

[Methods] There is a problem that the average and variance of hospitalization days differ greatly depending on the characteristics of the disease. Therefore, we used multi-step regression combining a logistic regression model using binomial distribution that predicts the occurrence of hospitalization for each disease and multiple regression that predicts the number of days of general hospitalization based on hospitalization risk values for eight diseases. Using a data of about 101,466 people(Age 35~69) that can be tracable for 8 years, we have built a model to predict the Total number of hospitalization days in the future 5 years. However, cases where the total hospitalization days exceeded 100 days were excluded because they were considered to be largely affected by unobservable factors. Diseases targeted in this study are malignant neoplasms, cerebrovascular disease, cardiovascular disease, diabetes, pancreatic disease, liver disease, hypertension disease and kidney disease.

[Results] As a result of evaluation by 5 cross validation methods, it was confirmed that hospitalization days in the future 5 years can be predicted with 4.1% as an average absolute error (s.d. 2.9). Moreover, in the first layer model that predicts the occurrence of hospitalization, it was confirmed that the discrimination performance by AUC was more than 0.7 or more in each disease(e.g. 0.8 for cardiovascular disease).

Poster viewing A

Wednesday 4 September 1.30pm

Comparison of two stage and one stage meta-analysis of joint longitudinal and time-to-event data

Maria Sudell, Catrin Tudur Smith, Ruwanthi Kolamunnage-Dona
University of Liverpool

Context: Meta-analysis of joint longitudinal and time-to-event data is a growing area of research. Recent publications have suggested methods for one or two stage meta-analysis of such data. However (for meta-analyses generally) under certain circumstances, results from one and two stage analyses can differ (an example being that one stage methods calculate exact likelihoods, whereas two stage methods calculate approximate likelihoods). Given the complex nature of joint longitudinal and time-to-event data, it is important to establish how results from one or two stage approaches could differ, the extent of any potential discrepancy, and the reasons for any discrepancies, whilst establishing guidelines for best practice.

Objective(s): This research will compare results of one and two stage meta-analyses of joint longitudinal and time-to-event Individual Participant Data (IPD) under a range of conditions, and give recommendations for future research.

Method(s): Multi-study joint longitudinal and time-to-event data will be simulated under a range of conditions including differing levels of heterogeneity and association structure. One and two stage meta-analyses of these simulated datasets will be conducted, and the results compared. Code to fit the models assessed will augment currently available joint modelling software packages in R. Discrepancies between results from one and two stage approaches will be identified and discussed, with reference to potential reasons for the discrepancies. A demonstration in a real multi-study hypertensive dataset will also be presented.

Results: Results and conclusions will be presented and discussed at conference. Guidelines for future meta-analyses of joint longitudinal and time-to-event IPD will be presented.

Poster viewing A

Wednesday 4 September 1.30pm

Estimation of Hospitalization Occurrence Using a Neural Network Based on Longitudinal Clinical Features

Kaoru Kobayashi, Hirofumi Kondo, Yasutaka Hasegawa, Shuntaro Yui, Hideyuki Ban
Hitachi, Ltd

Japanese Health Insurance Societies provide health services aimed at promoting health and preventing lifestyle-related diseases for their members. For effective health services, health condition of the members should be grasped for past, present and future based on healthcare data accumulated the society. Although we have developed prediction technique for future lifestyle-related disease onset and medical cost of that, prediction of hospital admission is also important because hospitalization has a large financial burden on the members. Therefore, in this study, we developed a predictive model for hospitalization by diabetes whose complication tends to be severe using longitudinal data of medical checkup and past medical history. It is difficult that applying well known time series analysis methods to the longitudinal data because the data is yearly data in other words the number of the data points observed is very small. Therefore, in our proposed method, firstly we generated longitudinal features which represent transition of health conditions from the longitudinal data. Secondly, in order to perform efficient and effective feature extraction from large dimensional explanatory variable vector consists of the longitudinal data and the longitudinal features, the vector is divided based on medical knowledge into several groups. And then, the divide vectors are compressed and hospitalization occurrence is predicted by multi modal neural network. The proposed method applied to the 60,000 people's longitudinal data which consists three years data for explanatory variables and five years data for objective variables. As a result, the AUC of the proposed method was 0.90 and the number of false positive reduced 12% from the GLM modeled using the longitudinal data. From the result of investigation, this is because that the proposed method is more accurate than the GLM especially for members whose health condition is worse non-linearly. In conclusion, it was shown that the transition of health condition contributed to the improving accuracy and the interpretability.

Poster viewing A

Wednesday 4 September 1.30pm

HealthyR Notebooks: Democratising open and reproducible data analysis

Riinu Ots
The University of Edinburgh

Introduction: An urgent priority for medical science is the empowerment of clinicians to robustly perform their own reproducible data analyses. Traditionally, data analysis has required complex and expensive software run on costly computer equipment. Furthermore, R is traditionally known for having a steep learning curve. Recent developments in the R ecosystem have made statistical programming more accessible, allowing us to efficiently teach R to non-statisticians.

Methods: “Data notebooks” combine analysis and reporting in one document to provide an intuitive interface between the researcher and their data analysis; this interface is the future of open and reusable data analysis. Our solution called “HealthyR Notebooks” harnesses the RStudio, R Markdown and the tidyverse packages, giving users access to state-of-the-art data science algorithms.

Results: We have modified our established and successful course, “HealthyR: R for Healthcare Data Analysis”, into a globally scalable, open access resource. This format not only provides the necessary training foundation but delivers participants a fully functioning toolbox with which they can immediately commence their own data analyses. The transition from traditional to literate programming has been well received by pilot attendees. Take-home message: R Notebooks are an intuitive and efficient way to teach and use R, the HealthyR Notebooks will be a freely available resource by the end of this year.

Poster viewing A

Wednesday 4 September 1.30pm

Creating an online mathematics and statistics learning community

Rachel Hilliam, Gaynor Arrowsmith, Alexander Siddons, Derek Goldrei, Cath Brown
The Open University

There is an increasing focus within Higher Education on the wider student experience. Student engagement and a well-developed community are associated with greater levels of retention student satisfaction and success. It is therefore crucial not only to support students at all stages of their student journey, but also to create a space where they can benefit from peer support and interact with the wider mathematics and statistics (M&S) community.The School of Mathematics and Statistics at The Open University has a long tradition of engaging students outside the ‘classroom’ environment and has needed to adapt how this support has been provided as the number of face-to-face opportunities has diminished. One such initiative was to provide a subject forum with the specific remit of offering advice on study planning – the module advice forum.The forum is hosted on an M&S Study website which provides dedicated resources to help inform module choice and study planning, which are generally not easy to access from pan-university websites. This includes lists of examination results and student satisfaction ratings on individual M&S modules. These have been collated over several years, so that students can see trends in order to help inform their decisions.Resources are also available for students to self-assess their readiness to start their next module, along with targeted advice about topics which they might need to revise or alternative modules to consider. Careers advice and guidance specific to M&S is also provided along with links to accreditation bodies such as the IMA and RSS.Discussions in the forum have flourished creating a true learning community. This space also provides students with the opportunity to engage more fully in issues such as curriculum development and delivery of student support leading to improvements in the structure of M&S qualifications, influencing the content of new modules, more effective assessment strategies, and better ways of supporting students.

Poster viewing A

Wednesday 4 September 1.30pm

Making the most of intensive longitudinal data

Tanja Krone
TNO

We want to know everything that happens all the time, especially when doing research. One way to adhere to this wish, is Ecological Measurement Assessment (EMA, also known as Experiences Sampling Measurements). Using EMA, people are asked to answer questions multiple times a day or week based on timepoints (e.g., every two hours) or on events (e.g., each meal), are asked to wear sensors (e.g. black carbon) or to measure variables themselves (e.g., glucose). This gives a rich dataset combining different kinds of measurements over time. However, the data is always less then perfect: up to 80% missed measurements occur, due to forgetfullness, drop-outs and faulty equipment. Furthermore, the designs are often based on convenience, allowing for collection on workdays, or whenever a participant 'feels like it'. Last but not least, several kinds of data-streams may be combined, such as heart rate and questionnaires, measured respectively every minute and three times a day. As such, EMA data poses an enormous challenge, both in seeing its possibilities, and working within its limits.We aim to create a protocol advising on how to collect and analyze EMA data, taking into account the different study designs, the preferred options of the applied scientist that needs to work with them, the used methods, combining different data-streams, and the handling of missing data. To this end, we will simulate datasets that are practically feasible to collect yet statistically optimal in the data-richness, and analyze them with several different techniques, combining practical research design with innovative statistical modelling. Our aim is to compare both the designs tested and the methods used, to find the optimal data-collection design and statistical analyses for the diverse range of EMA data. We will present the detailed plans on how we handled this, the designs and methods considered, and (hopefully) our first results.

Poster viewing A

Wednesday 4 September 1.30pm

Parallel Computation for Seeded MDS Based on K-means Clustering

Xiangyu Meng[1], Jian Huang[1], Finbarr O' Sullivan[1], Liam Ó Súilleabháin[2], Zhaoyan Xiu[1]
[1] *University College Cork,* [2] *Division of Research at Kaiser Permanente*

Parallel Computation for Seeded MDS Based on K-means ClusteringMultidimensional Scaling (MDS) is a generic name for a family of dimension reduction algorithms used to configure points in low dimensional space to maintain pairwise distance between them as accurately as possible. Although MDS is powerful for identifying the underline pattern of high dimensional data, the time and space complexities are at least quadratic in the number of points, N, which makes it computationally difficult or impossible for data sets when N is very large. To solve this problem, we developed a seeded MDS approach. We also examine the possibility of a seeded approach to use the parallel computation to speed up the computation of the proposed seeded MDS.Our approach consists of three steps:Step1. An MDS analysis is carried out to map n = K + 1 cases to a K-dimensional space as seeds. A seed case can be selected as cluster centres obtained by clustering analysis of the data.Step2. Using the seed values as a reference, an optimization procedure is developed to map each non-seed case so that its distances to the seed points in the K-dimensional space closely match its corresponding distances from these cases. Step3. A segmentation method is applied to the dataset. Then we use a multithread method with the segmented data in R to accelerate the optimization process by parallel computations. The accuracy of the proposed algorithm is evaluated based on the cMDS method by total distance difference - accumulating the distance of corresponding points, estimated by our approach and cMDS respectively in K-dimensional space. The algorithm is implemented in R and demonstrated on simulated and real datasets. The memory usage and computation time and accuracy as a function of the number of cases N will be reported. The relative performance of different approaches to the selection of seed cases will also be reported.

Poster viewing A

Wednesday 4 September 1.30pm

Reporting of planned power and statistical design in published digital education intervention studies for health professionals: a cross-sectional analysis

Ram Bajpai[1], Josip Car[2]
[1] *Research Institute for Primary Care & Health Sciences, Keele University,* [2] *Centre for Population Health Sciences (CePHaS), Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore*

Objective: The aim of this analysis was to assess the reporting of planned power and statistical design in published digital education intervention studies for health professionals.

Methods: We performed a systematic search of Medical Literature Analysis and Retrieval System Online, Excerpta Medica database, Cochrane Central Register of Controlled Trials (Cochrane Library), PsycINFO, Cumulative Index to Nursing and Allied Health Literature, Education Resources Information Centre, and Web of Science published between 2007 and 2016. A total of 874 intervention studies were identified and categorised into online-offline, mobile digital education, and simulation-based modalities for pre and post-registration health professions' education. Of these, 242 studies were randomly selected for methodological review and thematic analysis.

Results: Nearly two-thirds (63.2%) of studies published after 2011 and primarily from north America (47.9%). More than two-third of these studies were parallel design (88%), single centre-based (95%), and published in subject-specific journals (70.3%). Median number of study authors was 5 (interquartile range [IQR]: 4 to 7), and very few studies (6.6%) reported statistician as co-author. The median sample size was 72 (IQR: 12 to 120) participants. However, less than one-third (30.6%; n=74) of the studies were explained their formal power analysis steps. Of these, 72 studies reported type-I error and 68 studies reported desired power explicitly. The preferred choice of desired power in these studies was 80% (n=55). Only 10.3% of studies were adherent to the CONSORT reporting guidelines. Reporting of power analysis was significantly ($p<0.001$) associated with the adherent of CONSORT guidelines. Only 9% studies described their strategy to handle missing data in analysis. Nearly half (45.9%) of the studies reported p-value statement, and 59.5% of studies reported the statistical software used in their data analysis.

Conclusions: This analysis indicates the association between the reporting of power analysis and adherent to CONSORT guidelines.

Poster viewing A

Wednesday 4 September 1.30pm

Harnessing repeated measurements of predictor variables: a review of existing methods for clinical risk prediction

Lucy M. Bull, Jamie C. Sergeant, Mark Lunt, Glen Martin, Kimme Hyrich
The University of Manchester

Background: Clinical prediction models (CPMs) can predict the risk of health outcomes for individual patients. The majority of existing CPMs are limited to only harnessing cross-sectional patient information and are unable to harness longitudinal medical data available in electronic health records (EHRs). Incorporating repeated measurements into CPMs provides an opportunity to enhance their performance in practice as they inform how patients change over time and are less sensitive to measurement error (1, 2). Our aim was to systematically review the literature to understand and summarise existing approaches to harnessing repeated measurements of predictor variables in CPMs.

Methods: Medline, Embase, and Web of Science were searched for articles reporting the development of a multivariable CPM for individual-level prediction, and modelling repeated measurements of at least one predictor. Information was extracted on: the method, its specific aim, reported advantages and limitations, and software available to apply the method.

Results: The search revealed 217 relevant articles. Three methodological aims for harnessing longitudinal information were identified: enabling updated predictions over time; inferring a covariate value at a pre-specified time, and accounting for the effect of covariate change. Seven distinct methodological frameworks were identified. These frameworks range from time-dependent covariates in survival analysis, through to joint models, and machine learning algorithms.

Conclusions: Identified frameworks and the methods within them can vary by the aims addressed, assumptions made, and their use of longitudinal information. Their applicability will depend on the clinical question, regularity of follow-up visits, sample size, and variability amongst the longitudinal predictors.

References1. Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. PLOS Medicine. 2013;10(2):e1001381.2. Sweeting MJ, Barrett JK, Thompson SG, Wood AM. The use of repeated blood pressure measures for cardiovascular risk prediction: a comparison of statistical models in the ARIC study. Statistics in medicine. 2017;36:4514-28.

Winner of Best Poster at YSM 2019

Poster viewing B

Thursday 5 September 1.10pm

Bayessian Survival Analysis of Cervical Cancer using Weibull Regression Model

Serifat Folorunso
Department of Statistics, University of Ibadan

The modeling and analysis of lifetime for terminal diseases such as cancer is a significant aspect of statistical work. This study considered data from 440 cases of patients treated from cervical cancer between 2006 and 2015.This research is to apply a Bayesian modeling approaches for prediction of patient survival covering both statistical methodologies for Bayesian survival models on cervical cancer. The study achieves the following objectives: (i) to describe and implement a Bayesian Weibull model to predict patient survival; (ii) to observe the impact of censoring on fitting a Weibull regression model with covariates.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. Our results show that censoring has an effect on the performance of the Weibull models in that, 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. Our 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%.

Poster viewing B

Thursday 5 September 1.10pm

Data Ethics as an Ontological Problem

Thomas King

While the character of artificial ‘intelligence’ and the nature of ‘bias’ are regularly aired, ‘data ethics’ is taken for granted. This is compounded by the longstanding ethical basis of data at the datum level - do we not already have a ‘data ethics’? Digital data and statistical data are totally different ideas yet ‘data ethics’ is often used without defining what is meant by ‘data’. This term, ‘data ethics’, causes confusion but it is about data used in bulk, for which both statistical and computational perspectives have a bearing. Ethics is about opening questions not right answers, and the frame proposed will help us to specify our questions about use and impact of data.This work sets out a (2 x 2) frame for understanding the distinction from data aggregated from the individual datum. This includes the separation of interpretation from stochastic issues of measurement error, model selection, missing data and sampling. Both of these work at individual and bulk levels e.g. causes (may) correspond to correlations at the bulk level. Measurement of an attribute of an individual has corresponding construct validity and reliability. Consent by individuals has worked well for direct interventions but is incoherent for aspects of linkage to future data.Case studies are established as effective for stimulating ethical deliberation, so examples are presented to illustrate this frame: individual lives might be saved by a surgical or pharmaceutical treatment, but who are the people saved by public health interventions? Coverage of sensors for data collection to be used in prediction, allocation and training will lead to conservative inferences simply in response to sparser sampling. Service efficiency improvement is popularly synonymous with cuts, but should such concerns be used by the public to restrict such data usage? Such examples lead directly to consideration of what one ought to do and evaluation of the impact of data use on society.

Poster viewing B

Thursday 5 September 1.10pm

Forecast Error: The History of Political Betting in the United Kingdom

Timothy Martyn Hill
Barclays Corporate

The "Forecast Error" series of articles in Significance started examining election predictors in 2015. Each article considered many predictors, but each article covered just one election. In 2018 we started a new chapter in the "Forecast Error" series where we examine an individual class of predictor more closely across many elections. This presentation will cover political betting.

Political betting is frequently used as a predictor of elections. The United Kingdom has a history of political betting that predates the 20th century. We briefly review the pre-20th century wagers, then examine the Majorities betting market that existed before World War 2, then examine fixed-odds betting on elections that evolved as high-street bookmakers became normalised in the 1960s, then look at the 21st century with markets such as exchange betting, spread betting, and prediction markets.

Poster viewing B

Thursday 5 September 1.10pm

Finite-Sample Bias in Pseudo-Out-of-Sample Testing and Cross-Validation

Ian Hunt
Monash University

This paper analyses finite-sample bias in cross-validation estimates of expected prediction error. A significant risk of positive bias is identified against flexible models in the context of limited sample sizes --- this bias has practical implications for assessing curve-fitting models in finance and economics, for example when comparing regime-change models, or modern machine-learning techniques, with less flexible model-averaging methods. The bias is also positively related to the number of data-points left out for cross-validation purposes --- this has implications for traditional pseudo-out-of-sample Diebold-Mariano tests (using “hold-out” windows) because these tests rely on a special case of cross-validation that leaves out large chunks of data.First, I give analytic results for general leave-“n”-out cross-validation. Secondly, I render pseudo-out-of-sample tests as a special case of cross-validation. I conclude that the bias against flexible models is higher for pseudo-out-of-sample-tests than with leave-one-out and K-fold cross-validation.The analytic results add evidence to Diebold's conclusion that “[t]he hunch that pseudo-out-of-sample analysis is somehow the ‘only,’ or ‘best,’ or even necessarily a ‘good’ way to provide insurance against in-sample over-fitting in time-series model comparisons proves largely false" (Diebold, 2015). Plugging-in better estimates of over-fitting into the traditional DM-test statistic may help; better estimates in this context include those based on the expected optimism methodology (Efron, 2004) and cross-validation methods that leave out fewer data-points than pseudo-out-of-sample hold-out windows. This plug-in strategy could also ameliorate the problem of low power (Hansen and Timmerman, 2015) in pseudo-out-of-sample tests.I analyse both cross-sectional and time-series models. The time-series section adds finite-sample results to Hyndman-et-al (2018), who prove that leave-one-out cross-validation estimates of expected prediction error converge in probability to true expected prediction error for auto-correlated time-series if the estimated model adequately accounts for the true time-series dependence (pseudo-out-of-sample testing appears to rely on an analogous assumption).A simulation study and an application to forecasting equity market returns complete the paper.

Poster viewing B

Thursday 5 September 1.10pm

Modelling Categorical Data in Frequency Domain via Mutual Information Approach

Oluwaseun Wale-Orojo, Ajibola Soyonka, Olaniyi Olayiwola
FUNAAB

Phenomenon whose occurrences are described with count data in a multi-dimensional contingency table is common in medical and epidemiological studies. This study thus obtained the density function(s) of nonlinear vectors via mutual information approach and its measure of dependence from multivariate exponential power distribution which is a member of multivariate elliptical contoured family. The obtained model which accommodates for light and heavy tailed member distributions depending on the shape parameter were used to establish result for nonlinear vectors from multivariate Laplace, normal, uniform, students' t and Fisher distributions. Multivariate dependency of live birth on maternal age as the number of pregnancy increases irrespective of maternal height advantage as claimed from previous study, was established from the theoretical model obtained.

Poster viewing B

Thursday 5 September 1.10pm

Generalised additive models applied to tourism data

Caterina Constantinescu
The Data Lab, University of Edinburgh

In this talk I will discuss some of the functionality offered by R package `mgcv`. I will cover two generalised additive models (GAMs) built using this package, and applied to time series data measuring the monthly number of visitors at two historic sites in Edinburgh, Scotland: Craigmillar and Edinburgh Castles. These GAMs achieved a close description of the observed data, and largely respected the patterns of seasonality and other particularities that these time series exhibited.One of these models was devised to forecast future visitor numbers, and used the data to predict itself. A separate explanatory model was also constructed, and included various additional predictors (considered to be plausible drivers of tourism activity), e.g., temperature, hotel occupancy data, economic factors etc. Relevant Google search trends were also included in the prediction, so I will briefly discuss the related package, `gtrendsR`. Supporting functionality from packages `itsadug` and `mgcViz` will also be discussed, as these were helpful for generating visualisations for the relationship between each predictor included and the outcome, as well as generating confidence intervals around the predictions.Results included a linear relationship (as mgcv will not ‘force’ the fit of a wiggly line when this is not supported by the data) between hotel occupancy indicators and castle visitors, as well as the discovery of interesting non-linear patterns over time, allowed to vary by castle as well as visitor ticket types.

Poster viewing B

Thursday 5 September 1.10pm

kth Inflated Poisson Regression Modelling - EM Algorithmic Approach

Sameera Viswakula, Srinal Fernando
University of Colombo

This research focuses on modelling inflated count data using Poisson regression models. Poisson distribution is widely used to model count data. However, in many real-life applications, frequency of some of the counts is significantly higher than that is expected by the Poisson distribution (eg. zero, one, etc.). Therefore, inflated Poisson regression models are used for these data. Commonly used statistical software offers only zero or one inflated Poisson regression modelling mainly due to the complexity of the parameter estimation. In this research, EM algorithms are derived to model inflated count data using k-inflated Poisson regression models (k=0,1,2,…). A simulation study is conducted to assess the validity of estimates. Real-life data sets are used to illustrate the usage of the newly developed EM algorithm based R package for inflated count data.

Poster viewing B

Thursday 5 September 1.10pm

The Lindley-Weibull Power Series Class of Distributions and its Application to Life Time Data

Boikanyo Makubate[1], Broderick O. Oluyede[2], Thatayaone Moakofi[1]
[1] *Botswana International University Of Science and Technology,* [2] *Georgia Southern University*

A new generalized class of distributions called the Lindley- Weibull Power Series (LiWPS) class of distributions proposed and studied. The new class will be flexible, also with desirable propeties including hazard function that shows increasing, decreasing and bathtub shapes. This class of distributions generalizes the Lindley power series and the Weibull power series class of distributions respectively. A special model of the LiWPS class of distributions, the new Lindley Weibull Poisson (LiWP) is considered and some of of its mathematical properties are obtained. The LiWP distribution contains several new and well sub models including (Lindley Weibull Logarithmic (LiWL), Lindley Weibull Poisson (LiWP), Lindley Weibull Geometric (LiWG) and Lindley Weibull Binomial (LiWB)). Maximum likelihood estimation technique is used to estimate the model parameters followed by a Monte Carlo simulation study. Finally an application of the LiWP model to a real data set is presented to illustrate the usefulness of the proposed class of distributions. Goodness of fit tests are used to compare the fit of the LBXII distribution with its sub-models and other models for a given data set.

Poster viewing B

Thursday 5 September 1.10pm

Multi-stage models of cancer and disease

Anthony Webster
NDPH, University of Oxford

Since Armitage and Doll's publication of "a multi-stage theory of carcinogenesis" in 1954 [1], the multi-stage model has underpinned our conceptual understanding of cancer. In the last few years, researchers have started to apply the model to other diseases, such as Amyotrophic Lateral Sclerosis (Motor Neurone Disease), providing new insights into how the disease can arise and progress [2]. Here the multi-stage model is simplified and extended through a simple mathematical recipe for composing independent, sequential and cumulative models of carcinogenesis or other multi-stage failure models [3]. Relationships between different published cancer models are clarified, and derivations of new and existing results are simplified. Potential limitations of the model are discussed in the context of recent cancer research. The result is a simple framework for combining biologically-motivated models of each step in a disease's progression, providing a mathematical toolkit to study the failure of complex systems, biological or otherwise.

[1] P. Armitage and R. Doll, "The age distribution of cancer and a multi-stage theory of carcinogenesis," British Journal of Cancer, vol. 8, no. 1, pp. 1–12, 1954.

[2] A. Chio, L. Mazzini, S. D'Alfonso, et al. "The multistep hypothesis of ALS revisited The role of genetic mutations", Neurology, vol. 91, no. 7, pp. E635–E642, 2018.

[3] A.J. Webster "Multi-stage models for the failure of complex systems, cascading disasters, and the onset of disease", https://doi.org/10.1101/476242, in press, PLOS ONE, (2019).

Poster viewing B

Thursday 5 September 1.10pm

Spatial Analysis on Hepatitis in Punjab, Pakistan during 2010-14 and Interpolation of Disease Rates

Majida Jawad[1], Jawad Kadir[2]
[1] *University of the Punjab,* [2] *Lancaster University*

Increasing rate of Hepatitis cases is world wise concerns. Pakistan is one the highest rated countries in the world where not only hepatitis rate is so high but also increasing very rapidly. Therefore, a microscopic spatial analysis of the hepatitis infection that can provide scientific information for further intervention and disease control is needed. Tehsil wise percentage of infection cases recorded by the Punjab, Bureau of Statistics in Multiple Indicator Cluster survey were included in this study. Changing pattern of infection rates were observed by the used of spatial autocorrelation and interpolation tools. Moran's I plots and Hotspot analysis were conducted to identify the high risk areas. The finding of the analysis showed that the areas that are far away from the capital of the districts are more vulnerable as compare to the others. The analysis also pointed out that the areas belong to the southern part of the Punjab like Rajanpur and Hafizabad are the high-risk clusters.

Poster viewing B

Thursday 5 September 1.10pm

Constrained Empirical Bayes Estimation in Multiplicative Area-Level Models with Risk Analysis Under an Asymmetric Loss Function

Elaheh Torkashvand[1], Mohammad Jafari Jozani[1]
[1] *University of Notre Dame,* [2] *University of Manitoba*

Consider the problem of benchmarking small area estimates under multiplicative models with positive parameters where estimates are constrained to aggregate to direct estimates for the larger areas. Constrained (hierarchical) empirical Bayes estimators of positive small area parameters under many conventional loss functions can not be expressed in closed forms and/or might not even exist. Under the conventional squared error or weighted squared error loss functions benchmarked Bayes and empirical Bayes estimators can take negative values. In this paper, we propose a loss function that guarantees positive constrained estimates of small area parameters under multiplicative models. The proposed loss function penalizes underestimation of the unknown parameters of interest more than the squared error and the Kullback-Leibler (KL) loss functions. The hierarchical and hierarchical constrained empirical Bayes estimates of small area parameters and their corresponding risk functions under the new loss function are obtained. Also, the asymptotic approximation of risk and its second-order unbiased estimator are provided. We implement the Jackknife method in order to reduce the bias of estimators of risk. Finally, the performance of the proposed methods is investigated using simulation studies as well as a real data analysis for estimating asthma rate in subpopulations.

Poster viewing B

Thursday 5 September 1.10pm

Predictive regression with p-lags and order-q autoregressive predictors

Harshanie Jayetileke[1], You-Gan Wang[1], Min Zhu[2]
[1] *Queensland University of Technology,* [2] *University of Queensland*

This study considers the predictive regression models with p lagged predictors while the predictor xt itself is an AR(q) model. Akin to the standard predictive regression models, the ordinary least squares produces biased slope estimates. The existing methodology is applicable in the cases, p = q =1 via plug-in approach, or more generally, the augmentation approach introduced by Amihud, Hurvich and Wang (2010) in the case of p=q. We show that, the augmented regression method works in general, for any p and q orders. Extensive simulation studies indicate that, (i) the existing bias reduction methods become unsatisfactory when the order q is misspecified, in terms of bias and the mean squared errors; (ii) the correct identification of both p and q values are essential to successfully remove the bias; and (iii) the general augmentation method can reduce the bias satisfactorily and produces the smallest mean squared errors. The empirical application to stock market data from S&P 500 Index illustrate, a procedure to identify the qth order of the predictors such as dividend yield, earnings-to-price ratio, etc... and to obtain bias reduced estimates.

Poster viewing B

Thursday 5 September 1.10pm

Bayesian inference for GARCH Model for Cryptocurrency

Wantanee Poonvoralak
Chulalongkorn University, Sasin School of Management

Objective: The cryptocurrency financial market is known to be one of the most popular and volatile financial market in recent years. It is different from others financial markets due to that fact it can verifies transactions through its peer to peer network instead of using a trusted central party. Hence, numbers of GARCH models using statistical inferences have been used to examine the volatility effects of this currency. We propose the Bayesian inference for GARCH model for this currency. We aim to provide empirical results to help the BASEL committees to make more informed decisions about modelling in cryptocurrency and to whether it should be uses as part of the regulatory market risk management.

Methods: We used the most recent cryptocurrency with time period from 28th April 2013 to 21st March 2019 from the top 2 as listed on the 21st March 2019, they are the Bitcoin (BTC) and Ethereum (ETH) with market capitalization of $71,048,680,219 and $14,548,360,669 respectively. We firstly applied GARCH(1,1) to these data using statistical inferences and found that estimated kurtosis and persistence parameters suggested a requirement of heavy tail model distribution like the Student t-GARCH(1,1). This was not the case for Bayesian Markov Chain Monte Carlo (MCMC) approach.

Results: We found cryptocurrency data to be similar to other financial data such as the Foreign Exchange (FX) series in term of their volatility effects using GARCH modeling. Bayesian inference can help capture the persistence parameters and the MCMC sampling ensured finite kurtosis. Our results supported the decision that the cryptocurrency have implications on market risk management for investors and that the BASEL committees should consider it as part of regulatory risk management framework.

Poster viewing B

Thursday 5 September 1.10pm

A New Alternative to Cohen's Kappa Coefficient of Agreement

Tarald O. Kvalseth
University of Minnesota, USA

When two observers (raters) are independently classifying items or observations into the same set of mutually exclusive and exhaustive categories, some measure of agreement between the observers is often of interest. While a number of such measures have been proposed, the so-called Cohen's kappa, or simply kappa, has become by far the most frequently used measure in various fields of study. The popularity of kappa is based on its simplicity and appealing interpretation and the important fact that it corrects for the potential agreement expected by chance alone. When the various probabilities or proportions of agreement and disagreement between the two observers are represented in terms of a square contingency table, the chance agreement of kappa is based on statistical independence involving the marginal probability distributions. However, in spite of its popularity, Cohen's kappa suffers from some well-known limitations related to its dependence on the marginal distributions. In particular, kappa takes on unreasonably small values when the two marginal distributions are approximately symmetrical and highly uneven (non uniform). Alternative measures have been proposed and aimed at trying to overcome the kappa limitations. However, as argued in this present paper, the problems with Cohen's kappa as well as its proposed alternatives originate in their definitions of chance agreement (or disagreement) based on the marginal distributions. The present paper proposes a new kappa alternative that does incorporate chance agreement, but the chance agreement is not based on the overall marginal probability distributions. It is, however, based on independent chance classifications by the observers. The new measure does indeed overcome the limitations of Cohen's kappa and appears to have only desirable properties. A weighted form of this measure for the case when the classification categories are ordinal is briefly discussed. Also, statistical inference procedure for the new measure is outlined.

Poster viewing B

Thursday 5 September 1.10pm

classNB: a classifier based on the negative binomial distribution

Xiaoqiang Wang[1], Tianyuan Yao[1], Chunmao Huang[2], Heng Ge[1]
[1] *Shangdong University (Weihai), China,* [2] *Harbin Institute of Technology at Weihai*

Overdispersion is a widespread phenomenon in most count data sets. The negative binomial distribution is commonly adopted to fit the over-dispersed count data. In current work we devote to construct a toolkit, a R package called classNB for the classification problems based on the negative binomial distribution. In this package, we considered the mixture models, hidden Markov models, zero-inflated negative binomial models depending on the different data structures. In order to accelerate the process of parameter estimations, we develop an efficient Expectation Maximization (EM) algorithm which can avoid the typical numerical solution embedding in the EM algorithm. We further theoretically proved the convergence of the novel EM algorithm for the mixture models as well as the hidden Markov models. The simulation studies demonstrated that our methods effectively reduce the runtime of parameter estimations. Moreover, the proposed approaches also gain the same classification performance as the typical methods. The simulated data and the real data illustrate that the proposed toolkit, classNB has a good ability to classify the count data such as the next generation sequencing read count data, the earthquake data, air population data, etc.

Poster viewing B

Thursday 5 September 1.10pm

Analyzing diagnostic test data when sensitivity is not constant

Geoffrey Jones[1], Wes Johnson[2], Cord Heuer[1]
[1] *Massey University,* [2] *UC Irvine*

The analysis of data from diagnostic tests commonly makes the assumption that the sensitivity and specificity of the test(s) are constants. This is particularly important when there is no gold-standard test and latent class analysis has to be used to estimate the disease prevalence and test characteristics. For example the Hui-Walter model (Hui & Walter, 1980) uses data from two conditionally independent tests in two distinct populations to get an identifiable model for the six parameters – the assumptions of constant sensitivity and specificity in each population are crucial for achieving identifiability. In practice there are often scientific reasons for assuming that sensitivity, at least, will vary across populations. We present a Bayesian latent class analysis of data from a study of prevalences of leptospirosis in New Zealand beef herds to show how varying test sensitivity might be modelled and analyzed. The results suggest that relevant scientific information, incorporated not just into the priors but also into the model structure, can lead to a more appropriate analysis.

Hui, S.L., and Walter, S.D. (1980). Estimating the error rates of diagnostic tests. Biometrics, 36, 167-171.

Poster viewing B

Thursday 5 September 1.10pm

Application of Interrupted Time Series in evaluating PUMA - an intervention to improve early detection of in-patient deterioration in children's hospitals

Katie Taiyari[1], Kerry Hood[1], Robert Turbey[1], Devina Allen[2], Damian Ronald[3], on behalf of the PUMA Study Team
[1] Centre for Trial Research, Cardiff University, [2] School of Healthcare Sciences, Cardiff University, [3] University of Leicester

Background and objectives: PUMA is a quasi-experimental design study aimed at developing, implementing and evaluating the 'PUMA Programme' – an intervention to help children's hospitals improve their 'early warning system' and prevent avoidable mortality and morbidity among in-patients. Aggregate monthly outcomes were collected for four sites before (pre-), during (phase-in) and after (post-) implementation of the intervention. The common approach when analysing data from such design using ITS modelling is to exclude the phase-in data from the time series analysis, which is the period the intervention was implemented. However, such an approach is not appropriate in a scenario where there is a possibility that the performance may have been altered during the phase-in period. In such a scenario, excluding phase-in data from ITS analysis would not only reduce the power of the study but also reduce the chance of detecting the true change.

Methods: We examined three different approaches of evaluating ITS trajectory: (1) including the phase-in period in the analysis, (2) considering phase-in period as post-intervention, and (3) excluding the phase-in period. We compare Akaike information criterion (AIC) and Bayesian information criterion (BIC) from each model to choose the appropriate approach.

Results and Conclusions: Data from four UK-based paediatric hospitals in the PUMA study were collected from May 2015-October 2018, covering pre-intervention, phase-in and post-intervention periods. Data was analysed using ITS modelling. The data was the aggregate number of children in a given month who experienced at least one of following critical events: mortality, unplanned admission to PICU, unplanned admission to PHDU, cardiac arrest, and respiratory arrest. Data from each hospital were analysed separately using Interrupted Time Series model. All the results will be presented.

Poster viewing B

Thursday 5 September 1.10pm

Latent class models for use in a clinical setting using cohort data

Charlotte Watson[1], Hannah Lennon[2], Nophar Geifman[1], Andrew Renehan[1]
[1] *University of Manchester,* [2] *IARC*

Objectives: Life-course approaches to determine risk factor exposure are increasingly favoured over traditional 'once-only' epidemiological determination. One approach is latent class trajectory modelling (LCTM), which clusters individuals' changes in exposure over time, and offers a public health strategy to identify early divergent adverse trajectories as potential intervention targets. There is increasing use of LCTMs in mainstream epidemiology, but often with poor model description, and an over-reliance on the BIC metric for model selection. Here, we aimed to explore methods in deriving and validating LCTMs from multiple cohorts to ensure that they properly represent observed patterns across different populations.

Methods: We interrogated three cohorts, AARP (N: 321,827), PLCO (N: 147,488) and WHI (N: 151,363), with longitudinal BMI as the exposure and cancer incidence as the endpoint of interest. We extended our previously published work (https://bmjopen.bmj.com/content/8/7/e020683) to develop (i) a scoping framework for random-effect structures; (ii) testing multiple start points to obtain the global maximums; (iii) facilitating model choice with alternative metrics (APPA, Odds of Correct Classification, Entropy, Mismatch and DoF) to BIC. We developed visual model assessment tools, like convergence plots.

Results: We illustrate a number of examples where deviation from model assumptions yield very different classifications. We identified computational challenges for cubic models greater than 4 classes with up to 20 start points. After arriving as preferred models, we show that LCTM improve the performance characteristics of BMI exposure, compared with once only BMI measures, however, this improvement is clinically modest. LCTM might best identify specific sub-populations that have particularly high risk for cancer incidence.

Conclusions: The study highlights that model selection needs to be undertaken with care and not based solely determined by lowest BIC. We emphasise that multiple start points should be tested when using these models.

Poster viewing B

Thursday 5 September 1.10pm

Elicitation of prior beliefs for data lying in the simplex

Anastasia Frantsuzova, John Paul Gosling
University of Leeds

Right-stochastic matrices are used in the modelling of Markov processes (transition matrices) with a property that their elements are non-negative and each row sums to one. Similarly, data representing proportions of a whole can be unified into the compositional framework with similar structural properties. If we consider the problem of estimating these probabilities from a Bayesian standpoint, a natural choice of prior for the above multinomial likelihood is the conjugate Dirichlet family. There also stands a question of subjectivity of the prior and sensible ways to elicit such beliefs. We will present an overview of current elicitation methods for prior beliefs with data lying in the compositional simplex space in light of the Sheffield Elicitation Framework (SHELF). Additionally, we propose an additional step to the elicitation procedure, which aims to encapsulate dependencies between elicited marginal distributions and to allow the experts to refine or assure their beliefs by sensibly allocating weights to regions of the simplex.

Poster viewing B

Thursday 5 September 1.10pm

Reactor Embrittlement Curves – A Cracking Puzzle

Caroline Pyke
National Nuclear Laboratory

Light Water Reactors (LWRs), such as Sizewell B in the UK, contain a steel Reactor Pressure Vessel (RPV). The continued safe operation of LWRs, and their life extension, requires improved understanding of the behaviour of the toughness of these steels with increasing exposure to radiation - embrittlement. The condition of the RPV is monitored with surveillance samples, which are representative of the materials and irradiation conditions of the actual vessel. These are periodically withdrawn and their toughness tested using the Charpy impact test. For each steel a series of samples are tested at different temperatures and the relevant data recorded. The resulting data returns a sigmoidal curve for each set of surveillance samples. Empirical models of embrittlement focus on the shift of the curve at an individual point with increasing irradiation. However, this approach is limited as it relies on good data from unirradiated samplesThe purpose of this work was to determine whether the whole embrittlement curve could be established empirically using material composition data and irradiation parameters. The curves were estimated using non-linear least squares fits to describe the steel condition using three parameters. Ordinary linear least squares regression was then used to estimate each of these parameters. One of these parameters, the upper horizontal asymptote, otherwise referred to as the Upper Shelf Energy (USE), was particularly well predicted. The USE is expected to decrease with increased exposure to radiation. However, there can be considerable variability and/or bias in the test data and this phenomenon is not always observed. Indeed, often positive shifts in USE are observed from the raw data. From our models of USE we can estimate shift in USE when exposed to various levels of irradiation. Comparing these predictions with observed shift in USE returns some clear biases in the results.

Poster viewing B

Thursday 5 September 1.10pm

Fecal microbiota, Vitamin D, serum biomarkers of inflammation and diet: a complex interactive network influencing colorectal cancer risk and prognosis.

Pietro Belloni[1], Davide Serrano[2], Chiara Pozzi[3], Bruno Fosso[4], Silvia Guglietta[5], Harriet Johansson[2], Patrizia Gnagnarella[6], Debora Macis[2], Valentina Aristarco[2], Nicola Segata[7], Federica Corso[8], Federica Bellerba[8], Maria Rescigno[9], Sara Gandini[8]
[1] European Institute of Oncology, [2] Division of Cancer Prevention and Genetics, European Institute Of Oncology IRCCS, Milan, Italy, [3] Mucosal Immunology and Microbiota Unit, Humanitas Research Hospital, Milan, Italy, [4] Institute of Biomembranes and Bioenergetics, Consiglio Nazionale delle Ricerche, Bari, Italy, [5] Department of Microbiology and Immunology Medical University of South Carolina, USA, [6] Division of Epidemiology and Biostatistics, European Institute of Oncology IRCCS, Milan, Italy, [7] Department CIBIO, University of Trento, Trento, Italy, [8] Department of Experimental Oncology, European Institute of Oncology IRCCS, Milan, Italy, [9] Unit of Immunobiology of Dendritic Cells and Immunotherapy, European Institute of Oncology IRCCS, Milan, Italy

Background Microbiota and Vitamin D are involved in several pathologies including cancer.

Methods A prospective case-control study (34 CRC patients and 33 age match-controls) was designed to investigate microbiota composition, vitamin D indicators, circulating biomarkers of inflammation and lifestyle were analyzed.

Results Our 16S metagenomic sequencing confirmed a higher abundance of Fusobacterium nucleatum, Bacteroides fragilis, Parvimonas micra, Solobacterium moorei and Porphyromonas sp. in CRC patients. F. nucleatum showed a seasonal trend with greater abundance in winter when 25-hydroxyvitamin D (25OHD) is significantly lower (P=0.0002). In a multivariate model, F. nucleatum was found to be significantly associated with 25(OH)D (P=0.036) independently of CRC status and season. The markers of inflammation IL-6 and hs-reactive C protein negatively correlated with Vitamin D but positively with F. nucleatum (all P<0.05). Vitamin D receptor polymorphisms were found to correlate with CRC (FokI ff were more frequent among cases than controls, P-value=0.03) and with the abundance of microbiome members. Genera such as Parvimonas, Peptostreptococcus and Fusobacterium were significantly enriched in CRC patients consuming a diet poor in fatty fish (source of vitamin D) and rich in carbohydrates. Lastly, F. nucleatum was found to be significantly associated with time to relapse (P=0.03) and multivariate models including F. nucleatum, Vitamin D and Adiponectin identify CRC with an area under the ROC curve that patients ranged from 80-90% after cross-validation and reached similar values predicting recurrence.

Conclusions These parameters may set the bases for a diagnostic tool capable to predict the presence and recurrence of CRC.

Poster viewing B

Thursday 5 September 1.10pm

How clean is your takeaway?

Aura Popa
The Statistical Company

Consumers that are using food takeaway delivery apps like Just East, Deliveroo, Uber Eat do not see the Food Hygiene Rating of the restaurant, but the rating of other consumers. In October 2018, a BBC investigation found numerous zero rated takeaways on Just Eat, the UK's biggest online platform. In some major cities, more than half of the zero rated takeaways are on the platform.With the aim of comparing the Food Hygiene Rating Scheme versus the rating given by the customer to see if there is any relationship between the two, this poster will analyse the options of restaurants from three different takeaway delivery apps, looking into 2 wards of Kensington and Chelsea Borough, London. The two different areas, one of the poorest and one of the richest in UK according to the Index of Multiple Deprivation, are chosen with the purpose of testing the hypothesis if there are differences between the food hygiene and consumers' rating. The insights will be displayed in a visual format using infographics containing geographical heat maps through Spatial Analysis.Should the law of Food Hygiene Rating Scheme be changed and ask all the take away outlets to publish their score not on the door, but on these apps?

Poster viewing B

Thursday 5 September 1.10pm

A Bayesian approach to incorporate ecological species detection uncertainty into a rarity index

Jafet Belmont[1], Marian Scott[1], Claire Miller[1], Craig Wilkie[1], Tom August[2], Philip Taylor[2], Steve Brooks[3]
[1] *University of Glasgow,* [2] *Centre for Ecology and Hydrology,* [3] *The Natural History Museum*

Occupancy models at a community level have been explored widely in the literature. These models allow occupancy patterns to be estimated when the detection probability for the species within the community is less than one. Individual species responses may vary widely within a community, making the task of identifying the ecological processes of interest that drive the observed occupancy patterns difficult, especially for rare and elusive species.Several studies have developed methods to quantify the rarity of a species at the community level, but none has accounted for detectability bias. Leroy, B. et al. (2012) proposed a multiscale index of relative rarity (IRR) that allows species rarity to be compared regardless of the spatial scale, geographic area or taxonomic group. The index is computed by assigning a rarity weight to each species based on the observed occupancy and a rarity cut off point determined by the user.For this work, a 2-stage modelling approach was developed to estimate Leroy's rarity index in a Bayesian setting while accounting for detection bias. The index is computed at the first step as a derived quantity from the estimated occupancy by using a multispecies occupancy model. The second stage takes into account the first stage estimation uncertainty and models the responses as a function of site-level ecological covariates to evaluate which drivers affect the composition of rare species in a community. This approach was applied to a presence-absence dataset for Dragonfly occurrences in lakes across the UK. A selection of species-level covariates were used to estimate occupancy and detection probabilities. The estimated species occupancy was then used to compute the rarity index which was modelled in a second step to evaluate the effect that landscape connectivity and environmental stressor covariates have at different spatial scales.

Poster viewing B

Thursday 5 September 1.10pm

Modelling the effect of body mass on seabird survival: A Bayesian MCMC approach

Zhou Fang[1], David Elston[1], Francis Daunt[2], Kate Searle[2]
[1] *Biomathematics and Statistics Scotland,* [2] *Centre for Ecology & Hydrology*

Many species of seabirds are vulnerable to changes in their local environment. Such changes may directly cause mortality, but often instead are more subtle and affect the energy balance of birds (for example by making it more difficult to find food), thus causing changes in body mass which may alter survival probabilities. In order to obtain a clear idea of such indirect impacts, it is therefore important to determine the relationship between bird mass and over-wintering survival rates.Past studies in this area have used quite small data sets and simple methods. We analysed larger data sets consisting of mass measurements and capture-recapture or re-sighting records for four species – kittiwakes, puffins, guillemots and razorbills.These data sets are challenging to model for a variety of reasons. Birds have varying (and often quite small) probabilities of being recaptured or resighted, with the possibility of trap dependence effects and changing degrees of recapture effort. Mass measurements may be rare in frequency, or show different patterns of variability between birds, between years, and in terms of the day of the year each was measured. Birds may have different distributions of masses according to their sex, which is often not known. Survival is also impacted by other effects due to each year's environmental conditions and the bird's age.We defined a data augmented Bayesian MCMC model that attempted to capture these effects, combining a model for between- and within-year variation in bird mass with one for the survival of birds. This was fitted to each of the species, allowing us to obtain new estimates for the relationship of bird mass with bird survival.

Poster viewing B

Thursday 5 September 1.10pm

How varying CD4 criteria for treatment initiation was associated with survivability of HIV-patients?: retrospective analysis of health records from India

Ram Bajpai[1], Himanshu Chaturvedi[2]
[1] *Research Institute for Primary Care & Health Sciences, Keele University,* [2] *National Institute of Medical Statistics, Indian Council of Medical Research, New Delhi, India*

Background: HIV treatment and care services were scaled up in India at national level in 2007 with objective to increase HIV-care coverage. CD4-based criterion (or HIV clinical stage III/IV irrespective of CD4) was used for antiretroviral treatment (ART) initiation with increasing threshold in later years.

Objective: This paper aimed to evaluate the survival probabilities by varying CD4 criteria (CD4 ≤200 for the period 2007-08, CD4 ≤205 for the period 2009-11, and CD4 ≤350 from 2012 onwards) for ART among of HIV-positive patients.

Method: This retrospective analysis included 127,949 HIV-positive patients aged ≥15 years. All patients were registered for ART treatment between January 2007 to December 2011 and followed-up till December 2013 in Andhra Pradesh state, India. Mortality rate per 100 person-years was calculated. Kaplan-Meier and Cox-regression analysis were used to explore the association with all-cause mortality.

Results: Median age at the treatment initiation was 34 years (interquartile range [IQR]: 29 to 40), and 45.9% were female. Median CD4 count was 172 (IQR: 102 to 240) at the time of treatment initiation, and 19.3% of them had ≤100 CD4 count. Mortality rate for the period 2007-08 (CD4 ≤200) was 8.5/100 person-years compared to 6.4/100 person-years at risk for the period 2012 onwards (CD4 ≤350). Earlier thresholds for treatment initiation showed higher risk of mortality (CD4 ≤200 (2007-08), adjusted hazard ratio [aHR]: 1.86, 95% confidence interval [95%CI]: 1.68-2.07; CD4 ≤250 (2009-11), aHR: 1.67, 95%CI: 1.51-1.85) compared to the later threshold of CD4≤350 for treatment initiation from 2012 onwards.

Conclusions: Increasing CD4 threshold for treatment initiation over time was independently associated with the higher risk of mortality. The findings of this study provided critical insight into developing an effective and efficient HIV-treatment, care, and support program, and create a sustainable way to respond to the HIV epidemic in Andhra Pradesh, India.

Poster viewing B

Thursday 5 September 1.10pm

Modelling of long-term vegetation change in Scotland's native forests

Jacqueline Potts[1], Alison Hester[2]
[1] *Biomathematics and Statistics Scotland,* [2] *James Hutton Institute*

Forests play a key role in climate change mitigation, adaptation and delivery of a range of ecosystem services. This study examines the relationship between long-term changes in forest vegetation in Scotland's native forests and contemporaneous changes in climate, pollution and grazing. Plots, originally surveyed between 1958 and 1983, were resurveyed in 2007, with plant species composition being recorded. Four types of forest habitat (ash, acid oak-birch, base-rich oak-birch, and pine) were considered, both separately and as an overall sample of 'forest' plots. Changes in species richness were analysed using a generalised linear mixed model, and changes in species diversity and the percentage cover of major species groups and individual species were also analysed. A test of multivariate homogeneity was used to assess whether dispersion (variation between plots within habitat groupings) had changed. Statistical models were also used to explore the effect of drivers associated with climate (temperature, precipitation), pollution deposition (nitrogen, sulphur) and the density of herbivores on the vegetation data. All four habitat types showed compositional change. There was no evidence of homogenisation, with the opposite trend being seen in most habitat groups. The analyses identified statistically significant effects associated with climate, pollution and grazing, with deposition of reduced nitrogen, in particular, being associated with species compositional changes. Notable species changes seen between surveys include increases in pteridophytes (bracken) and declines in forb cover, and a doubling in frequency and cover of Fagus sylvatica (common beech).

Poster viewing B

Thursday 5 September 1.10pm

Introducing an adaptation to the MRSea package to allow for non-overlapping spatial ranges in a spatial interaction term; the case of the African Vultures.

Lindesay Scott-Hayward[1], Monique Mackenzie[1], Claudia Faustino[1], Ortwin Aschenborn[2]
[1] *University of St Andrews,* [2] *University of Namibia*

This poster presents a method for overcoming the issue of including a spatial interaction term for non-overlapping spatial regions. The data consist ofgeo-referenced locations from satellite tagged vultures, recorded between March 2015 and November 2018. To characterise habitat preference, Bernoulli generalised additive models were fitted to nine environmental covariates along with a spatial interaction term; s(x,y, year-month) to predict probability of presence of the vultures. The bivariate smooth of coordinates was based on the CReSS method (Complex REgion Spatial Smoother) with SALSA2D to determine the number and location of knots (the flexibility of the smooth). The standard implementation in the MRSea package for the interaction term uses the same candidate knot locations across each level of the interaction but the parameter estimates may vary. For this analysis, there were some year-month combinations whose spatial range was completely separate to others and this created an estimation issue. The new framework proposed for such situations allocates the flexibility of the spatio-temporal term to each factor level, in this case each year-month. Rather than setting out say 300 candidate knot locations, evenly spaced through the full geographical range of each bird (as incorporated in the MRSea package), candidate knots were placed throughout the geographical range of each year-month. Thus, each knot location was only relevant to a particular year-month and the parameters associated with these were only effective within that year-month. This new framework allowed model convergence and greatly reduced both the number of parameters estimated and the computational time.

Poster viewing B

Thursday 5 September 1.10pm

Electrophysiological Signatures of Cognitive Impairment in Parkinson's disease – A Machine Learning Approach

Aoife Sweeney[1], Barry Devereux[1], Charlie Ong[2], John McKinley[3], Seamus Kearney[3], Julia Foy[2], Brian Murphy[4], Bernadette McGuinness[13], Peter Passmore[13]
[1] Queen's University Belfast, [2] South Eastern Health and Social Care Trust, Northern Ireland, [3] Belfast Health and Social Care Trust, Northern Ireland, [4] Queen's University Belfast/Brainwavebank Ltd.

Dry-EEG offers an inexpensive, non-invasive and faster method to assess cognition in aging clinical groups. Cognitive impairment is prevalent in Parkinson's disease (PD), with 50% of patients developing dementia within 10 years (Williams-Gray et al., 2013). The presence of mild cognitive impairment (MCI), particularly visuospatial dysfunction, has been associated with increased dementia risk. EEG has previously been shown to have predictive value for patient outcomes in mild cognitive impairment associated with Alzheimer's disease (AD-MCI), yet there is little information on the electrophysiological correlates of PD-MCI. A 12-month follow-up study is being conducted. PD patients (n=60) and matched controls complete a comprehensive neuropsychological assessment and a battery of EEG tasks. These computerized cognitive tasks assess resting state activity, attention, language, memory and visuospatial cognitive domains. The PD patient cohort consists of both PD-NC (normal cognition) and PD-MCI patients. The primary outcomes for this study are the ability of dry-EEG to differentially discriminate between cognitively normal and cognitively impaired PD patients, and to detect changes in PD cognitive status over 12 months. We present power spectral analysis and event-related potential components which are associated with cognitive dysfunction in PD. Differences between groups are assessed using randomisation tests with FDR-corrected pairwise comparisons. Correlation analysis with validated clinical measures such as the MoCA and Hoehn and Yahr scores are performed. Machine learning algorithms such as the 'Elastic Net' (Zou & Hastie, 2005), random forests and k-nearest neighbour models are used to identify EEG features which most accurately discriminate between patient subgroups. Finally, the quality of the EEG data obtained (in terms of signal to noise ratio and the percentage of epochs rejected), plus patient acceptability of this technology are measured.

Poster viewing B

Thursday 5 September 1.10pm

Birth-death models for population genetics

Anastasia Ignatieva[12], Jotun Hein[1], Paul Jenkins[2]
[1] *University of Oxford,* [2] *University of Warwick*

Given a sample of genetic data obtained at the present, a question of interest is how to reconstruct a possible genealogy relating the individuals through common ancestors in the past. The coalescent is a standard model for genealogies in population genetics, but standard formulations assume a constant, or deterministically growing, population size. The dynamics of a population exhibiting exponential growth, such as viral populations, can more naturally be modelled as a birth-death process, which captures the stochastic variation in population size over time.The history of the population is then described by a birth-death tree; bifurcations in the tree correspond to births and leaves correspond to deaths of individuals. The genealogy of a sample of individuals alive at the present time can be obtained by deleting extinct and non-sampled lineages from the full tree. The stochastic process corresponding to this genealogy is termed the reconstructed process.The reconstructed process, viewed backwards in time, has an inhomogeneous pure-death formulation (with a time-dependent death rate). We discuss the properties of this process, and show how this formulation can be used to derive distributions characterising the genealogy of individuals sampled from the population.

Winner of Best Poster at RSC 2019

H

I

J

K

L

M

N

O

P

Q

R

S

T

V

W

X

Y

Z

ROYAL STATISTICAL SOCIETY

DATA | EVIDENCE | DECISIONS

tourism northernireland

Belfast visit Belfast

Royal Statistical Society
12 Errol Street, London EC1Y 8LX
020 7638 8998
conference@rss.org.uk

rss.org.uk