

# RSS International Conference Abstracts Booklet

**CARDIFF, WALES 3-6 September 2018**

**[www.rss.org.uk/conference2018](http://www.rss.org.uk/conference2018)**

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

## 1.1 Contributed - Medical: Prognosis and Prediction

Tuesday 4 September 9am - 10am

### Tailoring prediction models for use in new settings: Individual participant data meta-analysis for ranking model recalibration methods

Joie Ensor,<sup>1</sup> Emma Martin<sup>2</sup>, Kym Snell<sup>1</sup>, Thomas Debray<sup>3</sup>, Mamas Mamas<sup>1</sup>, Carl Moons<sup>3</sup>, Richard Riley<sup>1</sup>

<sup>1</sup>Keele University, <sup>2</sup>University of Leicester, <sup>3</sup>University Medical Center Utrecht

**Background:** The availability of individual participant data (IPD) from multiple sources allows the external validation of a prediction model across multiple settings and populations. When applying an existing prediction model in a new population it is likely that it will suffer from some over or under fitting, potentially causing poor predictive performance. However, rather than discarding the model outright, it may be possible to modify components of the model to improve its performance using model recalibration methods. Here, we consider how IPD meta-analysis methods can be used to compare and select the most appropriate recalibration method, or whether a completely new model is warranted in a particular setting.

**Methods:** We examine four methods for recalibrating an existing logistic prediction model in cardiovascular disease across multiple centres: (i) re-estimation of the intercept, (ii) adjustment of the linear predictor as a whole (calibration slope), (iii) adjustment of individual heterogeneous predictor effects, and finally (iv) re-estimation of all model parameters. We use multivariate IPD meta-analysis to jointly synthesise calibration and discrimination performance across centres for each of the methods. The most appropriate recalibration method can then be evaluated based on the joint probability of achieving a given model performance in a new setting, using this to rank recalibration methods.

**Results:** We present a new Stata package allowing estimation of the joint probability of achieving a set level of model performance in a new setting for each recalibration method, therefore easily identifying the method with the highest probability. We show that the best recalibration method is case specific and promote the use of recalibration as opposed to developing new models unnecessarily when the probability of improved performance through recalibration is high.

**Conclusions:** Multivariate meta-analysis allows quantification of the most appropriate recalibration methods to improve the performance of an existing prediction model in new settings.

## 1.1 Contributed - Medical: Prognosis and Prediction

Tuesday 4 September 9am - 10am

### **Investigating the impact of competing events on prognostic model outcomes: A simulation study**

Lucy Teece, Kym Snell, Richard Riley, Danielle van der Windt  
*Institute of Primary Care and Health Sciences, Keele University*

Prognostic models are used in primary care to predict an individual's future health outcomes, including risk of disease progression and the development of further complications. The statistical methodology used to develop these models is often naïve to the presence of competing events, these are events which may prevent or alter the probability of the outcome of interest from occurring. There has been relatively little research about the impact of competing events in relation to prediction model research. The way in which competing events may affect prediction model outcomes (if not appropriately accounted for) include: erroneous estimates of prognostic factor associations, inflated absolute risk estimates, mis-calibration, and inaccurate risk group allocation. A simulation study was undertaken to investigate the impact of competing events on these identified prediction model outcomes under different scenarios. The purpose of this analysis was to identify which circumstances the presence of competing events can alter the predictive ability of prognostic models, and thus provide guidance as to when it may be necessary to use competing risks statistical regression methods for model development. The simulation begins with an investigation into the amount of bias in non-parametric estimates of absolute risks when competing events are ignored. Multiple scenarios, in which the number and proportion of events of interest and competing events are altered, are presented. Flexible parametric regression models are presented to demonstrate levels of mis-calibration in prediction model estimates for differing scenarios. Finally, scenarios incorporating various levels of association between prognostic factor variables and both the events of interest and the competing events provide insight on bias in estimated prognostic factor associations, when competing events are neglected.

## 1.1 Contributed - Medical: Prognosis and Prediction

Tuesday 4 September 9am - 10am

### The usability of the three-zone diagnostic decisions

Natasa Keizar

*University of Ljubljana, Faculty of Medicine, Slovenia*

A biomarker can be used as a surrogate for the diagnosis (or prognosis) if it is acquired in a less expensive, less invasive or easier way than the gold standard. Before claiming the candidate a surrogate, it is crucial to quantify how well it replaces the gold standard. The measures of diagnostic accuracy are often computed on a sample of patients where both measurements (the candidate and the gold standard) are obtained. In the talk we consider candidate biomarkers with numeric values. We discuss the problems in detecting the optimal cut point for differentiating between the positively and negatively diagnosed patients. This is usually determined by the means of ROC curve. Since the optimal cut is often impossible to obtain, the notion of "gray zone" has emerged. It is represented by two points that form the three zones (sets) of patients: the likely positive, the likely negative and the still undecided. Few definitions of how to compute gray zone can be found in the literature (Coste et al 2003, Cannesson et al 2011). We discuss them with regard to the prevalence of the positive patients and explore (theoretically and with simulations) their usability in different settings: varying prevalence, sample size and the distributions of biomarker. The final objective is to make some recommendation about the use of gray zone in medical research.

## 1.2 Contributed - Official Statistics & Public Policy: Datasets, admin data, and collaboration across the GSS

Tuesday 4 September 9am - 10am

### Dataset families: making statistics easier to find and apply

Darren Barnes,<sup>1</sup> Bill Roberts<sup>2</sup>

<sup>1</sup>ONS, <sup>2</sup>Swirrl

In a huge collection of statistics, how do users find the right data to solve their problem? How do they know what the data means and how to combine or compare data from different sources? The Office for National Statistics is working with data technology providers Swirrl and stakeholders from across the Government Statistical Service to investigate a new approach to these important questions. The central idea is the "dataset family", a simple concept that can have a big impact. It refers to thinking of statistical datasets in terms of what they are about, rather than where they came from or which department produced them. The GSS produces a huge range of important statistics, used by a diverse audience for diverse reasons. Users have a hard time locating the statistics they need, even when they know these exist. Working with a group of producers and users of statistics on international trade, we have explored and documented the 'jobs to be done' by data users and what makes them hard. Underlying our candidate solution is the use of standards to enable interoperability and the web to enable access. The hard part is identifying what needs to be standardised and how, helping data producers to adhere to those standards, and exploiting the enhanced capacity for automation this brings to make the search for data easier. Datasets in the same 'family' can share dimensions, measures or classification codes that allow us to exploit the relations between them to assist users. This paper will present our progress so far in tackling these problems and the results of testing with prospective users. We are keen to share our thoughts about what we can do next and get as much experience and expertise from across the RSS to help us develop something really transformational.

## 1.2 Contributed - Official Statistics & Public Policy: Datasets, admin data, and collaboration across the GSS

Tuesday 4 September 9am - 10am

### Using administrative data to estimate rates of casualties resulting from road-traffic collisions

Rob Johnson,<sup>1</sup> Chris Jackson<sup>1</sup>, James Woodcock<sup>2</sup>, Anna Goodman<sup>3</sup>

<sup>1</sup>MRC BSU, <sup>2</sup>MRC Epidemiology Unit, <sup>3</sup>LSHTM,

We combine administrative data from multiple sources to estimate rates of road traffic injuries and deaths, and how these vary with demographic characteristics, road types and mode of transport. We use the STATS19 database, which records the number of reported road traffic injuries, in groups defined by the age, gender and mode of transport of the parties in the collision, and the road type, in each year from 2005 to 2015. The exposure of each group to the risk of injury is defined in terms of distance travelled per year. This is estimated by combining data from the National Travel Survey, the Driver and Vehicle Licensing Agency (DVLA), the Office for National Statistics, and the Department for Transport (DFT). Counts of injuries by group are modelled with a negative binomial regression, adjusting for distance travelled, demographic predictors, road and modes of transport and their interactions. This enables us to estimate the rate at which road users of a particular age, gender and mode of transport are injured or killed in a collision with another user of a particular age, gender and mode of transport, on a specific type of road. We discuss the statistical challenges, which arise mainly because not all relevant information is recorded in a single dataset. For example, the survey data are a biased sample of all road trips, underreporting commercial travel, and do not record the types of roads travelled on. This missing information is imputed based on the DFT data, which record distance travelled by road type but not demographic data, supplemented by DVLA data on goods vehicles. This model is part of a larger model to estimate the health impacts of travel policies and scenarios, particularly whether health gains from increased cycling and walking are offset by changes in injury rates.

## **1.2 Contributed - Official Statistics & Public Policy: Datasets, admin data, and collaboration across the GSS**

**Tuesday 4 September 9am - 10am**

### **Small Area Estimation of Fuel Poverty in England – collaboration between BEIS and ONS**

Katie Allison

*BEIS*

This presentation investigates improving methods for estimating sub-regional fuel poverty levels and is a follow up on from the presentation last year titled Small Area Estimation of Fuel Poverty in England (Katie Allison). It summarises progress from an ongoing collaboration between the Department for Business, Energy and Industrial Strategy (BEIS) and the Office for National Statistics (ONS) to develop alternative estimates of fuel poverty at a sub-regional level. The focus is on developing a multi-level small area estimation model with a binary survey response variable sourced from the English Housing Survey and a set of covariates that have been compiled from integrated administrative sources. The presentation will cover the following:- User needs for sub-regional fuel poverty estimates- Model selection- Uncertainty estimates- Results- Next steps for the collaboration Developing an alternative methodology ensure our official statistics remain relevant, robust, reliable and fit for purpose, and will allow BEIS to move away from current reliance on modelled Experian data. It also demonstrates the potential benefits of integrating survey and administrative data, collaborative working and the development of modelling expertise across the GSS.

### **1.3 Contributed - Applications of Statistics - Universal Applications**

**Tuesday 4 September 9am - 10am**

#### **Analysis of Linguistic Change Using Chain Graph Models for Structural Inference**

Craig Alexander, Ludger Evers, Tereza Neocleous, Jane Stuart-Smith  
*University of Glasgow*

Graphical models provide a visualisation of the conditional dependence structure between variables, making them an attractive inference tool. The improved readability makes this an appealing approach to represent complex model output to a non-technical audience. In this work, we introduce a novel approach using graphical models, namely a chain graph model structure. The model structure can be inferred from two parts: Directed edges between explanatory and response variables are modelled using a hierarchical model with multivariate response. Relationship between response variables is modelled using a Gaussian graphical model, using precision estimates obtained from the aforementioned hierarchical model. We present an application of this approach using linguistic data obtained from the Sounds of the City corpus, which consists of a real time corpus of Glaswegian speech. From this data, we look to recover the underlying model detailing which factors have contributed to vowel change throughout the century.

### **1.3 Contributed - Applications of Statistics - Universal Applications**

**Tuesday 4 September 9am - 10am**

#### **Massively parallel iterative Bayesian nonparametrics for cosmological parameter estimation**

Ben Moews, Joe Zuntz

*The University of Edinburgh*

Current efforts in cosmological parameter estimation often suffer from both the computational costs of approximating distributions in high-dimensional parameter spaces and the wide-spread need for model tuning. Specifically, calculating the likelihoods of parameter proposals through simulations imposes high computational costs, leading to excessive time requirements per experiment. We propose and implement an iterative approach that incorporates Bayesian nonparametrics, allowing for the more accurate drawing of samples from an approximated posterior distribution. Exploiting both parallelisation and recent advances in statistical methodology, we describe a novel method of massively parallel posterior approximation and provide the astronomy community, as well as the wider statistical community and practitioners from other application areas, with an easy-to-use and self-optimizing tool for accelerated parameter estimation.

### **1.3 Contributed - Applications of Statistics - Universal Applications**

**Tuesday 4 September 9am - 10am**

#### **Tracking data analytics as a tool for performance management of football players: A Case Study from an Elite Football Academy**

Varuna De Silva

*Loughborough University*

Football is a topic that millions of people in the world can relate to. The emergence of low-cost smart sensors have enabled continuous monitoring of humans under dynamic conditions, of which player tracking in sports is a popular application. In this exploratory study, we investigate the suitability of a widely available, yet under utilized data source (i.e. GPS based player tracking data) related to player performance, to inform management decision making in football. We try to answer various questions about player performance, such as positional demands, match day demands and preparation, and injury patterns of a group of football players from an elite football academy. While there are various types of data that are collected, GPS tracking data tends to be the data sources that has been consistently collected and available for the last 5 seasons. The following research questions are analyzed and discussed in the paper using hypothesis testing, ANOVA analysis, Regression analysis and pattern recognition. Is there a relationship between commonly used physical performance metrics in football? What are the positional demands in football training? How can we better train the players to mirror match performance? How could we track physical performance of players and relate it to injuries and subsequent recovery? We will discuss, using the results of the analysis, the practical use of statistics in elite sports, and how the insights are utilized to assist coaches and support staff make decisions.

## 1.4 Contributed - Social Statistics

Tuesday 4 September 9am - 10am

### Methods for analysing life history sequence data containing rare and uncommon states

Jennifer Prattle,<sup>1</sup> James Nazroo<sup>2</sup>, Bram Vanhoutte<sup>2</sup>

<sup>1</sup>*Department of Social Statistics, University of Manchester*, <sup>2</sup>*The Cathie Marsh Institute for Social Research, University of Manchester*

Sequence analysis methods are used to identify typical life course patterns from longitudinal data. Such patterns are experienced by a moderate to high number of people in the population. However, a small proportion of individuals may have experience of an uncommon event, such as homelessness, unemployment, illness or a prison spell. Investigating the influence, or predictors, of these experiences requires grouping life course sequences that contain a specific rare state into a distinct cluster or clusters, that collectively contain the majority of sequences with that state. Where multiple clusters exist for the same uncommon experience, there should be some distinguishing feature such as the timing or duration of the state. Identifying distinct clusters of rare state sequences from life course data is difficult when standard methods, such as optimal matching, are used. Optimal matching involves calculating pairwise dissimilarities between sequences using a series of insertion, deletion or substitution operations, each of which has an associated cost. A clustering algorithm is applied to the resulting dissimilarity matrix. A common approach is to set costs as constant irrespective of the state distribution or other sequence features. The dynamic Hamming method is a prevalent alternative but allows only substitution operations. Where these techniques are applied, rare state sequences are typically distributed across different clusters or allocated into one all-encompassing group of outliers. The use of two alternative methods - the period dependent Chi-square distance metric, and an optimal matching approach with state dependent insertion and deletion costs - is demonstrated and evaluated using life history employment data from the English Longitudinal Study of Ageing. The Chi-square distance metric calculated across the full time frame produces a theoretically viable model, comprised of distinct and interpretable clusters for each rare state of interest.

## 1.4 Contributed - Social Statistics

Tuesday 4 September 9am - 10am

### **Tackling Selection Bias in Sentencing Data Analysis: A New Approach Based on Mixture Models, Expert Elicitation Techniques, and Bayesian Statistics**

Jose Pina-Sánchez<sup>1</sup>, Sara Geneletti<sup>2</sup>, John Gosling<sup>1</sup>, Marco Doretti<sup>3</sup>

<sup>1</sup>*University of Leeds*, <sup>2</sup>*London School of Economics*, <sup>3</sup>*University of Perugia*

For reasons of methodological convenience statistical models analysing judicial decisions tend to focus on the duration of custodial sentences. These types of sentences are however quite rare (8% of the total in England and Wales), which generates a problem of selection bias, and raises questions about the external validity of much of the literature on the topic of sentencing. In this talk we present an original approach capable of modelling simultaneously custodial and non-custodial outcomes. This is achieved by using a Bayesian framework to estimate simultaneously: i) a scale of sentence severity based on Thurstone's pair comparisons retrieved from a sample of judges and magistrates, ii) an outcome model regressing the scale of severity on a set of case characteristics, and iii) a measurement model to reflect the uncertainty associated with the sampling error in the views taken from our sample of judges and magistrates. Using this approach for a sample of offences of assault sentenced at the Crown Court we can assess the magnitude of the selection bias associated with limiting the analysis to cases sentenced to custody only, which is non-negligible.

## **1.4 Contributed - Social Statistics**

**Tuesday 4 September 9am - 10am**

### **The Multi-Trait Multi-Error Approach to Estimating Measurement Error**

Alexandru Cernat

*University of Manchester*

Measurement error is a pervasive issue in surveys. One of the most common approaches used to measure and correct for systematic errors in this context is the Multi-Trait Multi-Method approach. Thus, it is possible to separate method, random error and “true” score using an experimental design that combines multiple traits (i.e. questions) with multiple methods (i.e. answer scales). As with other statistical approaches that tackle measurement error the results of this model are biased if any other types of systematic error (such as social desirability) are present. In this paper we present an extension of this model, which we name Multi-Trait Multi-Error, that manipulates multiple characteristics of the question format using a within factorial design. Thus, it is possible to estimate simultaneously: social desirability, acquiescence, method, random error and "true" score. We will illustrate how to implement the design and show initial results using measures of attitudes towards immigration in the 7th wave of the Understanding Society Innovation Sample.

## 1.5 Contributed - Methods & Theory: Dynamic and Markov Models

Tuesday 4 September 9am - 10am

### Robust and Sparse Dynamic Models

Paola Stolfi,<sup>1</sup> Mauro Bernardi<sup>2</sup>

<sup>1</sup>National Research Council of Italy, <sup>2</sup>University of Padova

In this paper we deal with sparse time-varying models, namely sparse multivariate models whose structure evolves over time. Sparse multivariate models are useful tools for analysing high-dimensional data. Despite their relevance in nowadays theoretical and applied statistics, these models mainly rely on independent and identically distributed Gaussian observations. The Gaussian assumption is relevant because of the interpretation of the variance-covariance matrix or its inverse as marginal or conditional undirected graphs, respectively. However, it is a quite restrictive and unrealistic assumption leading to poor estimates when data are strongly contaminated by the presence of outliers. Our major contributions consist to provide a robust and sparse estimator of time-varying conditional means and variance-covariance (or precision) matrices. Specifically, we handle time-varying structures by adding a semi-parametric kernel into the objective function that assigns higher weights to observations that are closer to the current one, extending the recent work of Zhou et al (2010). As concerns the robustification procedure, we propose to minimise the gamma-divergence introduced by Fujisawa and Eguchi (2008) as a robust alternative to the Kullback-Leibler divergence. The approach based on the gamma-divergence leads to a Majorization-Minimization (MM) algorithm for parameters estimation, and, as a further methodological contribution, we investigate its asymptotic properties. We apply the proposed methodology to functional Magnetic Resonance Imaging (fMRI) data, that are naturally highly contaminated and show complicated spatio-temporal correlation structure. The identification of brain network dynamics reveals correlation patterns displaying evident spatial shapes due to the underlying anatomical structure, and they are used to detect problem with mental health when an unbalance of these networks arises. In our application we compare the estimated brain network dynamics with those obtained with alternative non-robust dynamic models, across specific patients.

## 1.5 Contributed - Methods & Theory: Dynamic and Markov Models

Tuesday 4 September 9am - 10am

### Multi-state Event Analysis with Dynamic Longitudinal Covariates and Time-varying Coefficients

Chuoxin Ma, Jianxin Pan

*School of Mathematics, University of Manchester*

An important feature of cardiovascular disease (CVD) is that a growing number of hospitalization for worsen medical conditions is associated with a higher risk of CVD death. To address the association between non-fatal recurrent CVD events and cardiovascular deaths in a natural way, we propose to model the whole disease process in a multi-state event framework. That is, the evolution of CVD is modelled as a process with multiple states and transition to a new state is made when a new event (coronary heart disease events, myocardial infarction or death for example) is occurred. In addition to baseline covariates, time-dependent prognostic factor such as blood pressure, which is measured intermittently, is also included to better predict the risk of new CVD events. Since blood pressure is significantly influenced by prior CVD history as demonstrated in several medical studies, there is a strong argument for the existence of dynamic feedback mechanism in the disease evolution process. To address this issue, past event feedbacks are introduced as dynamic covariates, influencing the level of time-dependent longitudinal biomarker and hence further affecting the CVD event rate. The goal of this manuscript is to develop a multi-state event model with dynamic longitudinal covariate and semiparametric coefficients. The unknown time-varying coefficients are estimated via local partial likelihood, with one-step backfitting algorithm to improve computing efficiency. Consistency and asymptotic normality of the proposed one-step backfitting estimators are provided. Simulation study shows that our proposed model and estimation procedure perform well. We apply the model and estimation method to a data set from the Atherosclerosis Risk in Communities Study.

## 1.5 Contributed - Methods & Theory: Dynamic and Markov Models

Tuesday 4 September 9am - 10am

### Estimation of Viterbi path in Bayesian hidden Markov models

Kristi Kuljus,<sup>1</sup> Jüri Lember<sup>1</sup>, Dario Gasbarra<sup>2</sup>, Alexey Koloydenko<sup>3</sup>

<sup>1</sup>University of Tartu, <sup>2</sup>University of Helsinki, <sup>3</sup>Royal Holloway, University of London

The Viterbi path is the estimate of the underlying state path in hidden Markov models (HMMs), which has a maximum posterior probability (MAP). It is a common tool in HMM inference. For an HMM with given parameters, the Viterbi path can be easily found with the Viterbi algorithm. In the Bayesian framework the Viterbi algorithm is not applicable and several iterative methods can be used instead. We introduce a new EM-type algorithm for finding the MAP path and compare it with various other methods for finding the MAP path, including the variational Bayes approach and MCMC methods. To demonstrate the performance of different methods in the Bayesian segmentation context, we present an example with simulated data. The main focus is on non-stochastic iterative methods because these are computationally more efficient, and our results show that the best of those methods work at least as well or better than the best stochastic methods. Our results demonstrate that when the primary goal in HMM inference is segmentation, that is estimating the underlying state path, then it is more reasonable to perform segmentation directly by considering the transition and emission parameters as nuisance parameters. Furthermore, in the Bayesian segmentation context the choice of hyperparameters influences the solution crucially. Our simulation studies show that hyperparameters determine largely the nature of the problem, the properties of the solution and they also control the influence of data.

## 1.6 Contributed - Communicating Statistics: Learning statistics in applied contexts

Tuesday 4 September 9am - 10am

### Graphical principles for creating effective data visualizations

Mark Baillie, Alison Margolskee, Baldur Magnusson, Andrew Wright, Ivan-Toma Vranesic, Julie Jones, Marc Vandemeulebroecke  
Novartis

Objectives: The goal of this presentation is to illustrate Good Graphical Principles for statistical graphics. The presentation is related to a separate poster and handout of a Graphical Principles “Cheat Sheet”.

Methods: Effective visualizations communicate complex statistical and quantitative information to facilitate insight, understanding, and decision making. But how do you make an effective graph? Start with identifying the purpose (i.e. the question) the graph will address. Knowing the purpose will help to select the correct graph type for the data being displayed. Use proximity and alignment of graphical elements to facilitate comparison. Give consideration to axis scales and orientations to support interpretation. Use colour sparingly and only for a specific purpose. Use descriptive and legible axis labels and annotations to help convey the message.

Results: We present good graphical principles to consider when creating graphs, and give examples of applying these principles to some commonly used graphs. These principles are helpful in graphical data exploration and the production of graphics for communicating analysis results and conclusions.

Conclusions: Good Graphical Principles are useful for the creation of clear and impactful graphics.

*References:*[1] Cleveland, WS (1985). *The elements of graphing data*. New York: Chapman and Hall.[2] Duke S, Bancken F, Crowe B, Soukup M, Botsis T, Forshee R (2015). *Seeing is believing: Good graphic design principles for medical research*. *Statistics in Medicine* 34 (22), 3040-3059.[3] Krause A, OConnell M (editors, 2012): *A picture is worth a thousand tables*. New York: Springer. [4] Robbins NB (2013). *Creating more effective graphs*. Chart House. [5] Tufte ER (2001). *The visual display of quantitative information*, 2nd ed. Cheshire, CT: Graphics Press.[6] Tukey J (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley. [7] Wong DM (2013). *The Wall Street Journal Guide to Information Graphics: The Dos and Don'ts of Presenting Data, Facts, and Figures*. WW Norton.

## 1.6 Contributed - Communicating Statistics: Learning statistics in applied contexts

Tuesday 4 September 9am - 10am

### Teaching the concept of prior probability to medical students, doctors and other non-statisticians using 'P maps'

Huw Llewelyn

*Aberystwyth University*

There are two types of prior probability: Unconditional 'base rate' (but conditional on the feature(s) defining a universal set) and a conditional 'non base rate' prior probability conditional on the features of a universal set and at least one of its subsets. The difference can be clarified during teaching by using 'P maps' in conjunction with Venn diagrams[1]. The probability of the defining features of a universal set conditional on any of its subset features is 'one' but this is not the case for non base rate priors. P maps can clarify this and the nature of prior probabilities in clinical practice and how they can be used with and without the assumption of statistical independence. This can be compared with the role of priors when estimating parameters by using random sampling. P maps can be used to show how P values can be used to estimate the probability of 'idealistic' replication (when a study is described impeccably by an author and repeated impeccably by a reader) and how prior information can be incorporated into the evidence to estimate the probability of 'realistic' replication. They can also be used to represent how a single finding is linked to a list of possible diagnoses, each diagnosis having a conditional non-base-rate prior probability. The probabilities of the latter can be updated with new evidence by multiplying the ratios of a pair of prior probabilities (one of which belongs to the postulated diagnosis) with the ratio of a pair of corresponding likelihoods using a derivation of the extended form of Bayes rule. Examples will be provided using data from a study on the differential diagnosis of acute abdominal pain.

*References*Llewelyn H, Ang AH, Lewis K, Abdullah A. (2014) *Picturing probabilities*. In *The Oxford Handbook of Clinical Diagnosis*, 3rd edition. Oxford University Press, Oxford. p 618-621.  
<http://oxfordmedicine.com/view/10.1093/med/9780199679867.001.0001/med-9780199679867-chapter-13#med-9780199679867-chapter-13-div1-3>

## 1.6 Contributed - Communicating Statistics: Learning statistics in applied contexts

Tuesday 4 September 9am - 10am

### Revealed: What Every Statistician Should Know

Neil Spencer

*University of Hertfordshire*

At the end of this talk, the speaker and audience will have created a set of topics that all statisticians should know about and thus what should be contained in courses that train statisticians. The existing state of statistical training will be challenged, arguing that it is based on a bygone idea of the requirements of the workplace, whether that be academia or the private/public/third sectors. The commonly produced programme of “here are the classical techniques” with “here are a few exciting new techniques” does not give developing statisticians the diet that is required. Instead, it is asserted that statisticians must be trained to be flexible. They must be able to adapt known techniques to new situations, understand the issues involved when existing approaches are challenged and have the skills to cope when learning new procedures. Developing this flexibility should thus be a main aim of any statistical training programme but how can this be achieved? What are the essential parts of the statistical armoury that underpin this flexibility? What can be dispensed with? What supporting skills from other disciplines should be included? There are limits on how much can be contained in any training so what should be prioritised? We will invite the audience to decide which topics should be included in every statistician’s training. Topic suggestions made by the author, supplemented by proposals from the floor, will be pitted head-to-head in a series of votes. From these will emerge a list of what the room believes every statistician should know and be contained in training for statisticians. The sequencing of the head-to-head votes will use a technique which optimises the speed with which topics are ranked. The results of the ranking will be presented for discussion at the end of the talk and preserved on the author’s blog at <http://neilhspencer.wordpress.com/>, along with details of the vote sequencing method.

## 1.7 Contributed - Methods & Theory: Event Time Analysis

Tuesday 4 September 9am - 10am

### **A genetic algorithm of the maximum likelihood estimation for the parameters of Marshall-Olkin generalized exponential distribution**

Iklım Balay

*Ankara Yildirim Beyazıt University*

The Marshall-Olkin generalized exponential (MOGE) distribution can be used quite effectively for modelling life time data set; see Marshall & Olkin (1997) and Torabi et al. (2018). It is common to use the maximum likelihood (ML) method for the estimation of the MOGE distribution parameters. However, ML equations cannot be solved analytically. Therefore, iterative methods are used to overcome problems encountered in solving likelihood equations. In this study, we use the genetic algorithm (GA) approach to obtain the ML estimators of the parameters of the MOGE distribution. The main advantage of the GA approach is that it doesn't require any strict regulations as the traditional search techniques. We use an extensive Monte Carlo simulation study to compare the performance of the GA with the other iterative techniques, such as Newton Raphson (NR), Nelder Mead (NM) and iteratively reweighting algorithm (IRA). The simulation results show that the performance of the GA approach is the best among others in terms of bias, mean square error (MSE) and deficiency (Def) criteria. At the end of the study, we analyze a data set taken from the literature to demonstrate the performance of algorithms.

*References: Marshall, A. W., & Olkin, I. (1997). A new method for adding a parameter to a family of distributions with application to the exponential and Weibull families. Biometrika, 84(3), 641-652*  
*Torabi, H., Bagheri, F. L., & Mahmoudi, E. (2018). Estimation of parameters for the Marshall-Olkin generalized exponential distribution based on complete data. Mathematics and Computers in Simulation, 146, 177-185*

## 1.7 Contributed - Methods & Theory: Event Time Analysis

Tuesday 4 September 9am - 10am

### Causal Mediation Analysis with Sequentially Ordered Mediators Using Cox Proportional Hazards Model

Shu-Hsien Cho, Yen-Tsung Huang

*Institute of Statistical Science, Academia Sinica*

Causal mediation analysis aims to investigate the mechanism linking an exposure and an outcome. However, studies regarding mediation effects on survival outcomes are limited, particularly in multi-mediator settings. The existing multi-mediator analyses for survival outcomes are either under special model specification such as probit models and additive hazard models or assuming a rare outcome. Here we propose a novel multi-mediation analysis based on the widely used Cox proportional hazards model without the rare outcome assumption. We develop a methodology under the counterfactual framework to identify path-specific effects (PSEs) of the exposure on the outcome through the mediator(s), and derive the closed-form formula for PSEs on a transformed survival time. Moreover, we show that the convolution of an extreme value and Gaussian random variables converges to another gaussian provided that the variance of the original Gaussian gets large. Based on that, we further derive closed-form expressions for PSEs on survival probabilities. Asymptotic properties are established for both estimators. Extensive simulation is conducted to evaluate the finite sample performance of our proposed estimators and to compare with existing methods. The utility of the proposed method is illustrated in a genomic study of lung cancer survival.

## 1.7 Contributed - Methods & Theory: Event Time Analysis

Tuesday 4 September 9am - 10am

### Support Vector Estimation of Counting Process Intensities

Antonio Eleuteri

*Royal Liverpool & Broadgreen University Hospital - Department of Physics, University of Liverpool*

In many applications, individual life histories can be seen as stochastic processes moving between states in a discrete state space. The states correspond to the status of an individual, and transition between the states correspond to occurrences of the events under study. Quite often the object of the study is the intensity (or rate) at which events occur. It's also common that life histories are incomplete (censoring phenomenon.) Let  $(\Omega, \mathcal{F}, \mathbb{P})$  be a probability space, and  $\{\mathcal{F}_t, t \in [0, 1]\}$  a filtration on it. Let  $N(t)$  be a counting process adapted to  $\mathcal{F}_t$ . By the Doob-Meyer decomposition we can define the martingale  $M(t) = N(t) - \int_0^t \Lambda(s) ds$  where  $\Lambda(s)$  is the intensity process of  $N$ . This paper contributes to the study of Aalen's multiplicative intensity model, which assumes that  $\Lambda(t) = \alpha(t)Y(t)$ , where  $\alpha(t)$  is the intensity function (representing the transition intensity on the individual level) and  $Y(t)$  is an observable stochastic process (which measures the size of the risk set just before  $t$ .) Our objective is to provide an estimate of  $\alpha(t)$ . We address the problem as the solution of the linear operator equation  $A(t) = \int_0^t \alpha(u) du$ , where  $A(t)$  is the data term. More specifically, we consider the approximate equation  $P\alpha \approx \hat{A}$  where  $P$  is the integral operator, and  $\hat{A}$  is an approximation to  $A$ . We consider the approximation provided by the Nelson-Aalen estimator:  $\hat{A}(t) = \sum_{t_j \leq t} Y(t_j) - 1$ ,  $t_j \leq t$ . Solving an approximate operator equation is in general an ill-posed problem, so we formulate the solution as the constrained minimisation of a regularisation functional in a reproducing kernel Hilbert space  $H$ . We show how we can formulate the constraints to take into account the accuracy of the data term. We also show how both the intensity and cumulative intensity can be recovered at the same time in terms of the unique support vector expansion on the space  $H$ . Finally, we give illustrative numerical examples on simulated and real data.

## 1.8 Contributed - Applications of Statistics: Customers and Clients

Tuesday 4 September 9am - 10am

### Monitoring customer complaint data using survival analysis to track batch performance

Cameron Hogg, Graham Warren, Marco Wibbelmann  
*SPD Development Company Limited*

Problem: Consumer contact data provides invaluable data on how product is performing in the field; in this case home pregnancy and ovulation tests. The current methods for monitoring contact data analyses the cumulative number of complaints per million devices. However if we use individual batches data to create the product's overall average of complaints per million devices, this may lead to misleading results due to censored data. The censoring occurs as a result of batches being at different stages of their release to market. Therefore it would be advantageous to examine new approaches to analyse this data. Contact data is limited to the variables collected at point of contact. Contact date, total batch size and grouping variables (e.g. contact type and location). Contacts consist of product praise, enquiries and complaints; this analysis focused on complaints. This research investigated whether alternative approaches to analysis of consumer complaint data could yield richer information regarding in-field batch performance.

Solution: We utilised non-parametric survival analysis techniques to track batch complaint rates relative to the product overall using Kaplan-Meier survival curves and hazard ratios. For the product overall, data across different batches was pooled into one data set summarising number of complaints and number of devices at risk on each day relative to first complaint received as a proxy for release date. We investigated methods of monitoring batch performance. Thresholds have been created for flagging batches that have higher than permitted complaints than the acceptable range for that product. These thresholds use both the Kaplan-Meier curve and the hazard ratio. This novel approach appears to provide better resolution of the data than the previous analysis method.

## **1.8 Contributed - Applications of Statistics: Customers and Clients**

**Tuesday 4 September 9am - 10am**

### **Understanding our customer base through the use of unsupervised customer segmentation and supervised CART analysis**

Anselma Dobson-McKittrick

*Dwr Cymru Welsh Water*

Dŵr Cymru Welsh Water (DCWW) is a 'not-for-profit' company which provides water and sewerage services to over 3 million customers. Every single penny we make goes straight back into keeping bills down, looking after our water, and providing the best service we can. To ensure that we are delivering a customer service which provides a better, more personalised service, as recognised by Ofwat (Water Service Regulation Authority), the Data Science team has carried out Customer Segmentation. Customer Segmentation is the practice of dividing a customer base into groups (or segments) which are comprised of individuals that are similar to each other. By understanding the similar shared needs of each group, communication and services can be targeted and tailored to the relevant segments, and those customers that are struggling or in vulnerable circumstances can be identified and supported. Data sources both internal and external were obtained to help capture our customers' characteristics, situation and behaviours. A technique of unsupervised learning, cluster analysis, was performed on these data sources, as the variable of interest was unknown. The results from this technique were then used within a supervised technique, Classification and Regression Tree (CART), to visually explain the key attributes in each cluster and to provide a predictive model to assist in carrying out customer segmentation dynamically e.g. on new customers or when customers' characteristics or situation change.

## 1.8 Contributed - Applications of Statistics: Customers and Clients

Tuesday 4 September 9am - 10am

### Dynamic Prediction of Propensity to Purchase by Landmark Modelling

Ilan Fridman Rojas,<sup>1</sup> Aris Perperoglou<sup>2</sup>, Berthold Lausen<sup>2</sup>, Henrik Nordmark<sup>3</sup>

<sup>1</sup>*University of Essex and Profusion Media Ltd*, <sup>2</sup>*University of Essex, Department of Mathematical Sciences*, <sup>3</sup>*Profusion Media Ltd*

Recent developments in the analysis of time-to-event data have allowed for methodological and predictive improvements on a number of fronts relative to Cox models. In particular, the landmark modelling framework of Van Houwelingen (2007) allows for the inclusion of several crucial extensions to the Cox model: incorporation of time-varying effects, gross violation of the proportional hazards assumption, and prediction of conditional probabilities given known survival/no re-purchase up to given time points. In the current study we present a novel application of this landmarking methodology: the modelling of transactional data to predict propensity to purchase. This use case presents a number of challenges, including data sets of considerable size for which many current statistical models and tools no longer scale to, and recurrent events with high frequency and multiplicity, often with time-varying covariates and strongly time-dependent effects. The application of the landmarking approach to this domain yields challenges and results distinct to those faced in epidemiological use cases. We present the results of such an application to subsets of a data set from a retailer with ~2m customers and 7 years of collected transactional data. We compare a Cox model with the Andersen-Gill model to allow for time-varying covariates, and then further extend the analysis to two variations of landmark models, thereby extracting a measure of the time-varying effect of the covariates, and producing dynamic predictions of probability of re-purchase which condition on time elapsed since the last purchase. The resulting model for predicting propensity to purchase by landmark modelling is the first of its kind to the best of the authors' knowledge.

## 1.9 Contributed - Environmental & Spatial Statistics

Tuesday 4 September 9am - 10am

### **Ozone and particulate matter air quality in Bogota and Mexico City: a comparative study**

Eliane R. Rodrigues,<sup>1</sup> Biviana M. Suarez-Sierra<sup>2</sup>, Guadalupe Tzintzun<sup>3</sup>

<sup>1</sup>*Universidad Nacional Autonoma de Mexico (UNAM)*, <sup>2</sup>*Universidad Nacional de Colombia*,

<sup>3</sup>*Instituto Nacional de Ecologia y Cambio Climatico*

In this work we use non-homogeneous Poisson processes with a Weibull rate function to estimate ozone, PM10, and PM2.5 exceedances rates of environmental thresholds. Versions of the model with and without change-points are considered. Parameters present in the those versions are estimated using the Bayesian point of view via Markov chain Monte Carlo methods. The models are applied to ozone, PM10, and PM2.5 data from Bogota's (Colombia) and Mexico City's (Mexico) monitoring network. Based on the results we compare the air quality of the two cities regarding those three pollutants. The comparison is made in terms of the behaviour of the mean function of the Poisson process and consequently, in terms of the rate at which exceedances of relevant thresholds occur. Results corroborate the common belief that PM10 poses a more serious problem in Bogota than in Mexico City, whereas ozone affects more Mexico City. However, when we move to the case of PM2.5, results show that problem with this pollutant may be more serious in Mexico City than in Bogota. The models also capture the effects of changes produced by preventive measures implemented by the environmental authorities in both cities as well as anthropogenic events that produce either an increase or a decrease in a given pollutant's concentration. This is a joint work with Biviana M. Suarez-Sierra (of whose Ph.D. Dissertation this topic is part) and G. Tzintzun.

## 1.9 Contributed - Environmental & Spatial Statistics

Tuesday 4 September 9am - 10am

### Assigning ambient air pollutant exposures using Bayesian spatio-temporal models

Neil Wright, Katherine Newell, Christiana Kartsonaki, Hubert Lam  
*University of Oxford*

Research into the health effects of ambient air pollution requires estimates of exposure. However, measures of ambient air pollution may only be available at locations different from the locations of the individuals in the study. Spatio-temporal models of ambient air pollution can be used to predict pollutant levels across a geographical area and assign individual exposures. The China Kadoorie Biobank study is a large prospective cohort study in ten regions of China. We aim to predict pollutant levels at clinic locations in the Suzhou region, which can then be used as proxies for individual pollution exposure in analyses of health outcomes. Daily measurements of six pollutants are available between 2013 and 2015 from monitors situated across the region. Weather data is also available from monitors in the region, but their locations do not coincide with the locations of pollution monitors or clinics. We use a Bayesian spatio-temporal model for each pollutant, incorporating geographical covariates and spatial and temporal correlations. We perform approximate Bayesian inference using Integrated Nested Laplace Approximations (INLA), with the R-INLA package for R. In addition, we want to use weather data such as temperature, pressure, precipitation and wind speed as covariates in these models. We do this by using two stages of models. Firstly we use spatio-temporal models for each weather variable. Then, predictions of the weather variables at pollution monitor and clinic locations are added as covariates in the models for the pollutants. Uncertainty from the weather models needs to be propagated to the pollutant models. We consider different approaches to propagate uncertainty between the models and compare the results.

## 1.9 Contributed - Environmental & Spatial Statistics

Tuesday 4 September 9am - 10am

### Estimates of CO<sub>2</sub> Fluxes from the City of Cape Town through Bayesian Inverse Modelling

Alecia Nickless,<sup>1</sup> Peter Rayner<sup>2</sup>, Robert Scholes<sup>3</sup>, Birgit Erni<sup>4</sup>

<sup>1</sup>*Nuffield Department of Primary Care Health Sciences, University of Oxford*, <sup>2</sup>*University of Melbourne*, <sup>3</sup>*University of the Witwatersrand*, <sup>4</sup>*University of Cape Town*

We present the results of an atmospheric inversion study carried out for Cape Town. The method of Bayesian Inverse Modelling uses prior estimates of carbon dioxide fluxes from anthropogenic and natural sources and an atmospheric transport model to model the concentration of CO<sub>2</sub> at two measurement points located on the borders of the city. The prior estimates of the fluxes are refined by the inversion through reducing the misfit between the observed and modelled concentrations. The degree to which adjustments can be made to the fluxes are determined by prior uncertainties prescribed to the fluxes. We demonstrate the use of the Bayesian Inverse modelling technique to solve this highly underdetermined problem. The median uncertainty reductions of the aggregated weekly flux estimates for the whole domain of Cape Town was 28%, but reach as high as 50% for some weeks. At the individual pixel-level, uncertainty reductions of the total weekly flux reached up to 98%. In a given pixel, the estimates of the total fluxes were improved, but the inversion under the current framework was less successful at making corrections to the individual anthropogenic and natural fluxes. We show the sensitivity of the posterior solution to different components of the inversion framework, which include the structure and prior estimates used in the flux uncertainty covariance matrix, the products used to derive the prior fluxes, and the structure of the control vector. The control vector determines the spatial and temporal resolution at which the fluxes are solved. A limitation of this inversion framework is the prior natural fluxes and their uncertainties. The correlation length prescribed between uncertainties in the natural fluxes has a large impact on the posterior estimates, with a small change in the correlation length leading to an aggregated flux over the domain changing from a carbon sink to a carbon source.

## 2.1 Medical: Sample size for risk prediction & prognostic model research

Tuesday 4 September 10.10am - 11.30am

### Sample size formulae for developing a multivariable prediction model based on expected shrinkage

Richard Riley,<sup>1</sup> Kym Snell<sup>1</sup>, Joie Ensor<sup>1</sup>, Danielle Burke<sup>1</sup>, Karel Moons<sup>2</sup>, Gary Collins<sup>3</sup>  
<sup>1</sup>Keele University, <sup>2</sup>University Medical Centre Utrecht, <sup>3</sup>University of Oxford

Background: When designing a study to develop a new risk prediction model, researchers should ensure their sample size is adequate in terms of the number of participants (n) and events (E) relative to the number of predictor parameters (p) considered for inclusion in the model. Current sample size calculations are based on “rules of thumb”, such as at least 10 events per predictor parameter (EPP), which receive much debate and criticism.

Objectives: To produce a new sample size formula for studies developing a prediction model with either binary or time-to-event outcomes. Specifically, to identify in advance of data collection, the sample size needed to minimize the expected optimism in predictor effect estimates, and thus the expected shrinkage required after model development.

Methods: We derive a closed-form sample size formula, based on utilizing the heuristic uniform shrinkage factor of Van Houwelingen and Le Cessie. The formula allows researchers to identify n, p and EPP that correspond to an expected shrinkage factor close to 1, such as 0.9, that reflects low overfitting. It requires researchers to pre-specify the anticipated Cox-Snell of the model, and we show how to identify realistic values of based on published information (e.g. C statistic) for existing models in the same field. A suitable margin of error in other relevant estimates (e.g. overall risks) is also recommended.

Results: We illustrate the approach using examples of diagnostic and prognostic prediction models. This shows that, to target an expected shrinkage factor of 0.9, a new diagnostic model for Chagas disease requires an EPP of 3.9 and a new prognostic model for recurrent venous thromboembolism requires an EPP of 23.

Conclusion: Blanket rules of thumb for sample size are inappropriate, and our alternative proposal allows sample size and EPP to be tailored to the particular model and setting of interest.

## 2.1 Medical: Sample size for risk prediction & prognostic model research

Tuesday 4 September 10.10am - 11.30am

### Simulation-based sample size calculations for studies validating a prediction model

Kym Snell<sup>1</sup>, Joie Ensor<sup>1</sup>, Danielle Burke<sup>1</sup>, Laura Bonnett<sup>2</sup>, Bob Phillips<sup>3</sup>, Gary Collins<sup>4</sup>, Richard Riley<sup>1</sup>

<sup>1</sup>Keele University, <sup>2</sup>University of Liverpool, <sup>3</sup>University of York, <sup>4</sup>University of Oxford

Background: Sample size requirements for prediction modelling are often based on 'rules-of-thumb' such as requiring at least 100 (or even 200) events or non-events to validate a prediction model. Although often overlooked, it is not simply the performance measures that are of interest but also the precision in these estimates. Validation studies should therefore be large enough to estimate performance measures (such as calibration and discrimination) with reasonable precision.

Objectives: We propose a simulation-based approach for determining appropriate sample sizes when planning to externally validate prediction models.

Methods: By specifying the desired precision in predictive performance measures, and using the reported distribution of the linear predictor from the development cohort, a simulation approach can be applied, in which datasets of increasing size are generated until the confidence intervals for the predictive performance measures (C-statistic, calibration-in-the-large, calibration slope) are sufficiently narrow.

Results: This presentation will show sample size simulations for a variety of settings including those with varying degrees of mis-calibration, which may mimic data from a different population with a different prevalence level for example. The approach will also be illustrated using a real example.

Conclusion: In situations where the distribution of the linear predictor can be ascertained, our simulation-based approach allows tailored sample size calculations for external validation studies.

## **2.1 Medical: Sample size for risk prediction & prognostic model research**

**Tuesday 4 September 10.10am - 11.30am**

### **Rules of thumb for sample size calculations in prediction research**

Gary Collins

*University of Oxford*

Sample size considerations for developing prediction models, using logistic or Cox regression, are often based on the events-per-variable (EPV) rule of thumb. Simulation studies have recommended a range of values of EPV, from 5 up to 50. However, an EPV of 10 has become the de facto criteria, despite much debate and criticism. In this talk I will discuss discrepancies between these simulation studies, highlight some flaws, and show why any single rule-of-thumb based on the EPV is problematic. I will conclude that we need to move beyond EPV, and tailor the sample size to the setting and problem at hand.

## **2.4 Social Statistics: Capacity & capability priorities for complex analysis of large longitudinal studies**

**Tuesday 4 September 10.10am - 11.30am**

### **Understanding cross national regularities in the natural history of social inequalities in child development; unique insights from national birth cohort studies**

Liz Washbrook

*University of Bristol*

In this talk I will present findings from a long-running project on the developmental trajectories of children from different socio-economic groups that draws on longitudinal birth cohort data from the UK, the USA, Canada and Australia. The large social inequalities that continue to emerge in educational attainment and life chances are a source of great concern across the developed world. Effective public policy responses depend crucially on understanding when in childhood disadvantaged children are falling behind most rapidly. To what extent can inequalities in attainment be attributed to disparities that already exist when children enter the public schooling system? How far does schooling exacerbate or ameliorate these initial differences? Addressing these questions is difficult methodologically but essential if improvement efforts are to be targeted effectively and, as I will show, longitudinal cohort data are indispensable for this purpose.

## **2.4 Social Statistics: Capacity & capability priorities for complex analysis of large longitudinal studies**

**Tuesday 4 September 10.10am - 11.30am**

### **Skills to realise the full potential of UK longitudinal studies**

George Ploubidis

*UCL*

The UK has some of the most well-established and highly regarded longitudinal data collections in the world with the majority of these being multidisciplinary studies which integrate social and economic data with rich biological data, while a key area of recommendation of the ESRC's Longitudinal Studies Strategic Review was making the most of the potential for administrative record linkages in its longitudinal studies. I argue that in order to maximise the scientific potential of these remarkable studies and fully capitalise on their richness, a multidisciplinary approach to training and capacity building is needed. This could be in the form of a Q – step initiative for postgraduates that will allow the new generation of users of longitudinal studies to fully embrace the complexity of the new data landscape and master the methods needed for quantitative social science research in the 21st century.

## **2.4 Social Statistics: Capacity & capability priorities for complex analysis of large longitudinal studies**

**Tuesday 4 September 10.10am - 11.30am**

### **Evidence from ESRC Reviews of Longitudinal Studies and Skills**

Rebecca Fairbairn

*ESRC*

The International Review of the UK Longitudinal Studies portfolio was published by the ESRC in May 2018. The ESRC is currently developing work on the representativeness of studies including special populations, migration and attrition. Methodological innovation in terms of an administrative data spine is another feature of future work. Another ESRC review, of skills and capabilities in social science is also accumulating evidence and some clear needs are emerging. Identified areas of need include quantitative and data skills, but the Longitudinal studies Review also identified limited understanding of the impacts of such analytical work. There is also a need for better understanding of how complex analysis translates into policy.

## **2.5 Methods & Theory: Multilevel models: imputing level-2 missing data. From theory to practice**

**Tuesday 4 September 10.10am - 11.30am**

### **Bayesian Multilevel Latent Class Models for the Multiple Imputation of Nested Categorical Data**

Daide Vidotto

*Tilburg University*

With this talk, I am going to introduce the use of Bayesian multilevel latent class (BMLC) model for the multiple imputation of nested categorical data. One of the advantages of the BMLC model is its flexibility, which enables it to automatically deal with possible interactions in the joint distribution of the dataset at hand. Furthermore, in the presence of variables observed at the higher level, the model allows for simultaneous imputations at both levels of the hierarchy. After a formal introduction of the model, the results of a simulation study will be presented, in which the performance of the BMLC is compared with the listwise deletion and JOMO methods.

## **2.5 Methods & Theory: Multilevel models: imputing level-2 missing data. From theory to practice**

**Tuesday 4 September 10.10am - 11.30am**

### **Multiple imputation of missing level 2 covariates in a multilevel model: analysis of the relationship between student ratings and teacher beliefs and practices**

Leonardo Grilli<sup>1</sup>, Maria Francesca Marino<sup>1</sup>, Omar Paccagnella<sup>2</sup>, Carla Rampichini<sup>1</sup>

<sup>1</sup>*University of Florence*, <sup>2</sup>*University of Padua*

We analyse the relationship between student ratings of university courses and several characteristics of the student, the course and the teacher. We exploit a survey collecting information about teacher beliefs and practices at the University of Padua in a.y. 2012/13. Student ratings are nested into classes, calling for multilevel modelling. However, due to survey non-response, about half of the teachers have missing values on practices (10 binary items) and beliefs (20 ordinal items on a seven-point scale). The problem is challenging due to the high percentage of missing values and the large number of categorical variables involved. We handle missing values through multiple imputation by chained equations, exploiting information at all hierarchical levels (level 2 covariates and summaries of level 1 covariates). The proposed approach turns out to be effective. We discuss the solutions adopted in the implementation of the approach, pointing out findings and proposals in the recent literature.

## **2.5 Methods & Theory: Multilevel models: imputing level-2 missing data. From theory to practice**

**Tuesday 4 September 10.10am - 11.30am**

### **Aggregating level-1 information for imputing level-2 missing data**

Omar Paccagnella<sup>1</sup>, Roberta Varriale<sup>2</sup>

<sup>1</sup>*University of Padua, Department of Statistical Sciences*, <sup>2</sup>*Italian National Statistical Institute - ISTAT*

One important issue imputing missing data at level-2 in multilevel research is the role played by the variables at level-1 to include in the imputation process. Strategies using either their manifest and/or latent group means have been investigated, as well as the comparisons across different imputation approaches. In this work we aim at enriching the literature on this topic proposing an extensive simulation study to investigate and compare the behaviour of different ways to aggregate level-1 information, that is using group means, medians, standard deviations or inter-quartile ranges of the level-1 data. Several conditions vary in our study: the type of the level-2 variables to impute (both continuous and binary), as well as the type of the level-1 variable to aggregate (both continuous and binary); the inclusion (or not) of the empirical Bayes estimates in the imputation process; the value of the Intraclass Correlation Coefficient (low or high) of the original multilevel data; the imputation approaches we adopted (Fully Conditional Specification or Predictive Mean Matching). Results highlight the advantages of including the empirical Bayes estimates in the imputation procedure. There are also some interesting findings looking at the imputed data that lie on the left tail of the distribution of the variable to be imputed.

## **2.8 Industry & Finance: Optimality issues in experimental design for business and industry**

**Tuesday 4 September 10.10am - 11.30am**

### **NCIS (New Constrained Imputed Solution) for Incomplete Designs**

Trevor Duguid Farrant

*Mondelez*

Mondelēz is the largest snacking company in the world with iconic brands Cadbury, Toblerone, Milka and many, many more. A substantial amount of money is spent on consumer studies and alternatives are sort after. Incomplete Designs (ICDs) are desired for consumer trials to reduce fatigue and costs but there are associated risks. In a Complete Design (CD) all consumers rate all products say 15 products 300 consumers that's a lot of products/consumers/questions to make/recruit/collect. In an ICD, each consumer sees a different subgroup of the products. If each consumer sees 14 products there is very little cost/time saving so this is not worthwhile. If each consumer sees 2 products there is a huge cost/time saving but a massive loss in the richness of the data, so again not worthwhile. The questions are 'Where is the sweet spot?' and 'What is the minimum number of consumers would you need before an ICD is too incomplete?' In conjunction with ICDs, NCIS has been developed to impute the missing scores for the study based on all consumers, all products and the individual consumer's scoring style. This allows the creation of a CD from an ICD for analysis purposes, useful for analyses, e.g., clustering, that only work on CDs.

## **2.8 Industry & Finance: Optimality issues in experimental design for business and industry**

**Tuesday 4 September 10.10am - 11.30am**

### **Blocking in multi-stage experiments**

R. A. Bailey

*University of St Andrews*

In a multi-stage experiment, the same experimental units are used in each stage but different treatment factors are applied at different stages. Constraints on processing imply that these units must be partitioned into blocks (such as batches or lots) at each stage. However, unlike in the classical situation, the blocks are not inherent, and the designer of the experiment can choose the partition into blocks at each stage. Is it better to align the Stage 2 blocks with the Stage 1 blocks as far as possible or to make them as orthogonal to each other as possible? In either case, how should treatments be assigned? In the simplest case, the treatment factors applied at each stage can be orthogonal to the blocks in that stage. In other cases, there may be one or more stages in which the treatment factor(s) applied in that stage must have each level applied to whole blocks. Both of these are comparatively straightforward compared to the case where there is one (or more) stage(s) in which the allocation of the treatments to experimental units must be that of an incomplete-block design. In the talk, I will develop some general principles for good design, along with methods for evaluating competing designs.

## **2.8 Industry & Finance: Optimality issues in experimental design for business and industry**

**Tuesday 4 September 10.10am - 11.30am**

### **Designing industrial experiments with functional independent variables with pharmaceutical applications**

David Woods

*University of Southampton*

In this talk, some novel methodology will be presented for the optimal design of experiments when at least one independent variable is a function (e.g. of time) and can be varied continuously during a single run of the experiment. Hence, finding a design becomes a question of choosing functions to define this variation for each run in the experiment. The work is motivated by, and applied to, experiments in the pharmaceutical industry.

## **2.9 Data Science: The future of SPC in a big data world**

**Tuesday 4 September 10.10am - 11.30am**

### **Data Science and Predictive Analytics at BT**

Blaise Egan

*BT*

BT is using predictive analytics in a number of areas across the company. Apart from driving activities typically associated with the area, for instance marketing, we also use predictive analytics in our operations to improve efficiencies or pre-empt faults. The application of analytics is typically constrained by operational environments that require bespoke implementations that limit the techniques that can be applied. I'll be presenting an example from the business and also briefly outline our efforts in establishing a best practice analytics framework.

## **2.9 Data Science: The future of SPC in a big data world**

**Tuesday 4 September 10.10am - 11.30am**

### **The impact of Big Data on creating insightful SPC**

Gillian Groom

*Minitab Ltd*

Statistical process control charts, were created in the 1920s by Walter Shewhart to monitor and control the variation in the quality of components and finished products. He developed a visual tool that was easy to produce and interpret. These were created at a time when measurements of parts had to be taken manually, and there were no computers and all the charts and calculations had to be completed by hand. Thus sampling was essential. The objective was to separate common cause variation (or random variation) from special (or assignable) causes. Today businesses still need to separate common cause variation from special causes, however nowadays complete inspection and hence measurement of all parts is not uncommon, and we now have powerful computer tools to produce our charts and separate the common from the special cause variation. The question is can the control charts still deliver and how can we use this data to understand the underlying trends. In this presentation we will be: Reviewing, false alarm rates and the impact on a control chart of using the complete data. Studying the impact of at sub-group size on control chart results Evaluating the performance of control charts based on full v sampled data Making recommendations for Insightful SPC with “big data”

## 2.9 Data Science: The future of SPC in a big data world

Tuesday 4 September 10.10am - 11.30am

**“Does SPC have a role anymore?”**

Omar McCarthy

*AstraZeneca*

“Does SPC have a role anymore?” Walter A. Shewhart developed the concept of a control chart for monitoring process inputs or outputs. Many organisations are well versed in the application and have been using it for decades, so is there still a need for this practise? Is it still confined to manufacturing? Is it a declining practise? “Does SPC have a role anymore?” This talk will cover: a brief overview of statistical process control the purpose of SPC and it's history an evaluation of SPC and whether there is still a need for it the application of SPC in non-manufacturing areas concluding argument for a greater and broader sustained use of SPC

## **2.10 : Exploring the gender data gap**

**Tuesday 4 September 10.10am - 11.30am**

Deirdre Appel

*Open Data Watch*

The fundamental principle of the 2030 development agenda is to leave no one behind. Achieving real inclusion - and monitoring progress - will require a significant improvement in the availability of data, specifically data disaggregated by sex. But what gender data are available for the measuring and monitoring of the Sustainable Development Goals and where do the gender data gaps lie? This session will explore the role of gender data in SDGs and present new research from Open Data Watch and Data2X's project, Mapping Gender Data Gaps.

## **2.10 : Exploring the gender data gap**

**Tuesday 4 September 10.10am - 11.30am**

Claudia Wells

*Office for National Statistics*

Gender issues remain a topic of fierce interest and debate. Here at ONS we already publish data on the gender pay gap, harassment, domestic violence and suicides. We want to present a fully coherent picture to help policy makers and answer the questions that matter most to the public. We need to ask ourselves, do we have the right data, does it come from the right people and is it collected in the right ways?

How inclusive are our current statistics?

Who is left out and “invisible”?

Can we use new sources of data and data science techniques to improve our understanding?

This presentation will give an overview of ONS’ work to address the gender data gap including the launch of a Centre for Inequalities and the development of international measures of gender equality through our work on the Sustainable Development Goals.

## **2.10 : Exploring the gender data gap**

**Tuesday 4 September 10.10am - 11.30am**

Natasha Davies

*Chwarae Teg*

An overview of how Chwarae Teg, Wales' leading gender equality charity, utilises gender data and the challenges of accessibility and availability.

## **Keynote 2 - Statistical Inference for Analysis of Massive Health Data: Challenges and Opportunities**

**Tuesday 4 September 12 noon – 1pm**

Xihong Lin

*Chair, Department of Biostatistics, Harvard T.H. Chan School of Public Health*

Massive data from genome, exposome, and phenome are becoming available at an increasing rate with no apparent end in sight. Examples include Whole Genome Sequencing data, large-scale remote-sensing satellite air pollution data, digital phenotyping, and Electronic Medical Records. The emerging field of Health Data Science presents statisticians with many exciting research and training opportunities and challenges. Success in health data science requires strong statistical inference integrated with computer science, information science and domain science. Examples include signal detection, network analysis, integrative analysis of different types and sources of data, and incorporation of domain knowledge in health data science method development. In this talk, I discuss some of such challenges and opportunities, and illustrate them using high-dimensional testing of dense and sparse signals for whole genome sequencing analysis, integrative analysis of different types and sources of data using causal mediation analysis, and analysis of multiple phenotypes (pleiotropy) using biobanks and Electronic Medical Records (EMRs).

### **3.1 Medical: Flexible hazard regression models for time-to-event data**

**Tuesday 4 September 2:10pm - 3:30pm**

#### **Smooth additive survival models and selection: if a tree falls in the forest, why?**

Simon Wood, Nicole Augustin  
*University of Bristol*

I will discuss methods for estimating smooth additive mixed survival models using reduced rank smoothing splines in a proportional hazards setting. I will focus particularly on smoothing parameter selection, using an empirical Bayes approach based on Laplace Approximate Marginal (partial) Likelihood. Model selection will also be discussed, both in the presence of random effects (frailties), and when there are moderately large numbers of effects to screen. The methods will be illustrated with an example of modelling tree survival for natural and managed forests, as part of long term forest health monitoring.

### **3.1 Medical: Flexible hazard regression models for time-to-event data**

**Tuesday 4 September 2:10pm - 3:30pm**

#### **Multidimensional penalised splines for hazard and excess hazard regression**

Mathieu Fauvernier

*University of Lyon*

Taking advantage of a recent theoretical framework for general smooth models (Wood et al. 2016), we present a penalized approach for hazard and excess hazard models in time-to-event analyses. Baseline hazard and functional forms of covariates are specified using penalized natural cubic splines with associated quadratic penalties. Interactions between continuous covariates and time-dependent effects are dealt with by forming a tensor product smooth. The smoothing parameters are estimated by either optimizing the Laplace approximate marginal likelihood criterion (LAMML) or likelihood cross-validation criterion (LCV). The regression parameters are estimated by direct maximization of the penalized likelihood of the survival model, therefore avoiding data augmentation and Poisson likelihood approach. The approach will be illustrated with examples based on real data.

*Reference: Wood, S. N., Pya, N. and Säfken, B. (2016), "Smoothing parameter and model selection for general smooth models," Journal of the American Statistical Association, 111(516), 1548-1563.*

### **3.1 Medical: Flexible hazard regression models for time-to-event data**

**Tuesday 4 September 2:10pm - 3:30pm**

#### **Flexible Bayesian Excess Hazard models using Low-Rank Thin Plate splines**

Manuela Quaresma

*London School of Hygiene & Tropical Medicine*

Regression models for the excess hazard became the preferred modelling tool for cancer survival research using population-based data. In this setting, the model is formulated as the additive decomposition of the total hazard into two components: the excess hazard due to the cancer and the population hazard due to all other causes of death. We introduce a flexible Bayesian regression model for the log-excess hazard where the baseline log-excess hazard, and any non-linear and non-proportional effects of covariates are modelled using Low Rank Thin Plate splines. The model also accommodates random effects and hierarchical data structures. Using this type of splines will ensure that the log-likelihood function retains tractability and thus numerical integration is not required when evaluating the likelihood function. We also demonstrate how to obtain posterior inferences for the excess hazard and for net survival, a populational level measure of the survival only due to the cancer, after accounting for other causes of death. We illustrate this model using survival data for all patients diagnosed with colon cancer that live in London and were treated in a London hospital.

### **3.4 Social Statistics: Gendered trends: women, men and tracing trajectories of change**

**Tuesday 4 September 2.10pm - 3.30pm**

#### **The best places in Britain to be a woman: using local authority level data to generate an index**

Julia Griggs, Allison Dunachik  
*NatCen Social Research*

In 2017, the National Centre for Social Research conducted a rapid study for BBC Woman's Hour to identify the best (and worst) places in Britain to be a woman. The study combined a series of indicators related to women's quality of life into an index, and then used existing data to rank all British local authorities on that index. This presentation - the third and final one in the Gendered Trends session - describes the project's methodology, provides an overview of results, and discusses the challenges and limitations in using administrative data broken down by gender.

### **3.4 Social Statistics: Gendered trends: women, men and tracing trajectories of change**

**Tuesday 4 September 2.10pm - 3.30pm**

#### **Trends in mental health and self-harm: Adult Psychiatric Morbidity Survey 1993-2014**

Sally McManus

*NatCen Social Research*

There has been a threefold increase in rates of reported self-harm in young women since 2000. This second talk in a session on Gendered Trends presents analyses of data from the Adult Psychiatric Morbidity Survey: the world's longest standing mental health survey series to using consistent methods. After some years of stability, the gap between young women and men has widened for some indicators of mental illness and related behaviours, and narrowed for others. We describe the nature of these changes and discuss theories for its emergence.

### **3.4 Social Statistics: Gendered trends: women, men and tracing trajectories of change**

**Tuesday 4 September 2.10pm - 3.30pm**

#### **Public perceptions of gender issues: British Social Attitudes 1984-2017**

Eleanor Attar Taylor

*National Centre for Social Research*

This is the first of three presentations that draw together recent studies using survey and administrative data to examine gendered trends in Britain. Analyses of British Social Attitudes survey data will be presented, showing how the series can be used to map changes since the early 1980s in public attitudes to women, men, and gender roles. New results from the latest British Social Attitudes survey on attitudes to online sexist bullying, and unsolicited comments on a woman's appearance, will also be presented.

### **3.5 Methods & Theory: Understanding when selection bias can occur**

**Tuesday 4 September 2.10pm - 3.30pm**

#### **Causal Inference under Outcome Dependent Sampling**

Vanessa Didelez

*Leibniz Institute for Prevention Research and Epidemiology - BIPS*

I will review when and how it is possible to draw causal conclusions under outcome dependent sampling, e.g. case-control designs; some results are valid for more general situations where by design or accident sampling depends on the outcome. The focus is on "identifiability": does the available data, at least in principle (for 'very large' samples), allow us to consistently estimate the desired causal quantity? If the answer is 'no' then this is typically due to structural bias, i.e. to fundamental problems of design and available information. In case-control studies, we face the following potential sources of structural bias regarding causal inference: (1) Case-control studies are necessarily observational, so confounding is likely to be present. (2) Case-control studies are retrospective with sampling being conditional on disease status which means there is also a threat of selection bias. (3) A consequence of the retrospective sampling is that methods which depend on, or are sensitive to, the marginal distribution of the outcome cannot be used without some modification, since the required information is not generally available. This is potentially relevant to certain methods of adjusting for confounding as well as to the identifiability of typical causal effect measures, such as the average causal effect. While confounding is a problem of any observational study and has been widely addressed in the causal inference literature, points (2) and (3) are more specific to outcome dependent sampling and will be the focus here. I will specifically consider: identification of the null-hypothesis of no causal effect which is closely related to the non-parametric identification of causal odds ratios; further I will address how certain structural knowledge can enable us to reconstruct the full joint distribution; finally I will briefly compare these approaches with those that rely on additional knowledge of the population prevalence, such as standardisation, propensity scores, and instrumental variables.

### **3.5 Methods & Theory: Understanding when selection bias can occur**

**Tuesday 4 September 2.10pm - 3.30pm**

#### **Survivor bias in Mendelian randomisation analyses**

Stijn Vansteelandt,<sup>1</sup> Eric Tchetgen Tchetgen<sup>2</sup>, Stefan Walter<sup>3</sup>

<sup>1</sup>*Ghent University, and the London School of Hygiene and Tropical Medicine*, <sup>2</sup>*The Wharton School, University of Pennsylvania, PA*, <sup>3</sup>*University of California, San Francisco, CA*

Mendelian randomization studies commonly focus on elderly populations. This makes the instrumental variables analysis of such studies sensitive to survivor bias, a type of selection bias. A particular concern is that the instrumental variable conditions, even when valid for the source population, may be violated for the selective population of individuals who survive the onset of the study. This is potentially very damaging because Mendelian randomization studies are known to be sensitive to bias due to even minor violations of the instrumental variable conditions. In this talk, I will give insight into this problem of survivor bias and how it might affect Mendelian randomisation analyses. I will moreover discuss strategies for eliminating such bias. In particular, I will show that, interestingly, the instrumental variable conditions continue to hold within certain risk sets of individuals who are still alive at a given age when the instrument and unmeasured confounders exert additive effects on the exposure, and moreover, the exposure and unmeasured confounders exert additive effects on the hazard of death. I will then exploit this property to derive a two-stage instrumental variable estimator for the effect of exposure on mortality, which is insulated against the above described selection bias under these additivity assumptions.

### **3.6 Communicating Statistics: Data Journalism**

**Tuesday 4 September 2.10pm - 3.30pm**

#### **How the BBC personalises data stories**

Clara Guibourg

*BBC News*

The problem: Bringing dry data journalism full of numbers to life. The solution: Bring the reader into the story, make it personal and show how it's relevant to them. Data journalism is a great source of exclusive news lines - but also makes it possible to give the reader an answer to the question "What does this situation look like for me specifically?" The BBC Visual Journalism team has been creating personal calculators to do this for years, from 2013's Great British class calculator to more recent calculators on everything from house prices, tampon tax and footballers' wages.

### **3.6 Communicating Statistics: Data Journalism**

**Tuesday 4 September 2.10pm - 3.30pm**

#### **Turning a data avalanche into the regional daily's splash**

Claire Miller

*Reach*

The key to getting to grips with data journalism is remembering it is just like all other types of journalism...What is the story? Using data as the basis of stories isn't new, crime stats, exam results and elections have always made up part of the local news output, but with vast quantities of data now easily accessible and more scope to use online story-telling techniques, the Reach Data Unit has been working to combine functionality, originality and innovation in order to make data stories an everyday part of the news agenda.

### **3.6 Communicating Statistics: Data Journalism**

**Tuesday 4 September 2.10pm - 3.30pm**

#### **Simpson's Paradox and the adolescence of data journalism**

John Burn-Murdoch

*Financial Times*

As data journalism gains popularity and access to statistical tools becomes easier, are data journalists at risk of putting the cart before the horse? Using an example from the German election, this talk will explore the pros and cons of doing data science on tight deadlines, and ask whether data journalists need to spend more time on the qualitative side of their work before ploughing ahead in quantitative complexity.

## 4.1 Contributed - Medical: Meta-analysis

Tuesday 4 September 3.40pm - 4.40pm

### **Multilevel network meta-regression for population adjustment based on individual and aggregate level data**

David Phillippo, Sofia Dias, Tony Ades, Nicky Welton  
*University of Bristol*

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. We can relax this assumption if individual patient data (IPD) are available from all studies by fitting an IPD meta-regression. However, in many cases IPD are only available from a subset of studies. In the simplest scenario, IPD are available for an AB study but only AgD for an AC study. Methods such as Matching Adjusted Indirect Comparison (MAIC) create a population-adjusted indirect comparison between treatments B and C. However, the resulting comparison is only valid in the AC population without additional assumptions, and the methods cannot be extended to larger treatment networks. Meta-regression-based approaches can be used in larger networks. However, these typically fit the same model at both the individual and aggregate level which incurs aggregation bias. We propose a general method for synthesising evidence from individual and aggregate data in networks of all sizes, Multilevel Network Meta-Regression, extending the standard NMA framework. An individual-level regression model is defined, and aggregate study data are fitted by integrating this model over the covariate distributions of the respective studies. Since integration is often complex or even intractable, we take a flexible numerical approach using Quasi-Monte Carlo integration, allowing for easy implementation regardless of model form or complexity. Correlation structures between covariates are accounted for using copulae. We illustrate the method using an example and compare the results to those obtained using current methods. Where heterogeneity may be explained by imbalance in effect modifiers between studies we achieve similar fit to a random effects NMA, but uncertainty is substantially reduced, and the model is more interpretable. Crucially for decision making, comparisons may be provided in any target population with a given covariate distribution.

## 4.1 Contributed - Medical: Meta-analysis

Tuesday 4 September 3.40pm - 4.40pm

### **Multivariate meta-analysis of correlated outcomes: how much borrowing of strength do we expect?**

Miriam Hattle, Danielle Burke, Richard Riley  
*Keele University,*

A multivariate meta-analysis allows the joint synthesis of multiple correlated outcomes. Compared to a standard univariate meta-analysis of each outcome independently, the utilisation of correlation can lead to more precise summary results from a multivariate meta-analysis. However, the multivariate approach is more complex and not routinely adopted by researchers. Recently, a borrowing of strength (BoS) statistic has been proposed by Jackson et al. (2017), which quantifies the gain in information from using a multivariate meta-analysis over a univariate meta-analysis. It is calculated as the percentage reduction in the variance of the summary estimate between the multivariate and the univariate meta-analyses. A large BoS value indicates when there is larger potential for important differences between multivariate and univariate results. For example, in a meta-analysis of the effect of perioperative Ketamine for acute postoperative pain on nausea, the BoS was large (51.4%) and the univariate and multivariate meta-analysis odd ratios were very different, 0.68 (95% C.I.: 0.44 to 1.03) and 0.66 (95% C.I.: 0.49 to 0.88) respectively. Therefore, researchers would benefit from knowing how much BoS to expect in their meta-analysis dataset, to help decide whether a multivariate approach is potentially worthwhile. In this presentation, we will use empirical evidence from 43 meta-analyses to investigate meta-analysis characteristics (including the number of studies, magnitude of correlation, and amount of missing outcomes) for their influence on the magnitude of BoS. Furthermore, we will show mathematically that BoS is bounded by the percentage of missing data for the outcome of interest, under certain conditions. Recommendations will be provided.

Jackson D, White IR, Price M, et al. Borrowing of strength and study weights in multivariate and network meta-analysis. *Stat Methods Med Res* 2017; 26: 2853–2868.

#### 4.1 Contributed - Medical: Meta-analysis

Tuesday 4 September 3.40pm - 4.40pm

##### **Model-based network meta-analysis for time-course relationships: A union of two methodologies**

Hugo Pedder,<sup>1</sup> Sofia Dias<sup>1</sup>, Meg Bennetts<sup>2</sup>, Martin Boucher<sup>2</sup>, Nicky Welton<sup>1</sup>

<sup>1</sup>University of Bristol, <sup>2</sup>Pfizer Ltd

Background: Model-based meta-analysis (MBMA) is a technique increasingly used in drug development for synthesising results from multiple studies, allowing pooling of information on treatment, dose-response and time-course characteristics, which are often non-linear. Such analyses are used in drug development to inform future trial designs. Network meta-analysis (NMA) is used in Health Technology Appraisals (HTA) and by reimbursement agencies for simultaneously comparing effects of multiple treatments. Recently, a framework for dose-response model-based network meta-analysis (MBNMA) has been proposed that draws strengths from both MBMA and NMA.

Objectives/Methods: We aim to expand the MBNMA framework for modelling of time-course functions that allows for the inclusion of multiple study time-points using a Bayesian approach. Our methodology preserves randomisation by aggregating within-study relative effects and, by modelling consistency equations on time-course parameters, it allows for testing of inconsistency between direct and indirect evidence. Residual correlation between observations can be accounted for using a multivariate likelihood. We demonstrate this modelling framework using an illustrative dataset of 24 trials investigating treatments for pain in osteoarthritis.

Results: Of the time-course functions that we explored in our dataset, an Emax function allowed for the greatest degree of flexibility, both in the time-course shape and parameter specification. Our final model was a good fit to the data (posterior mean residual deviance of 291.4; 345 data points), and all treatments had a greater maximal effect than placebo. Treatment estimates were robust to the choice of likelihood (univariate/multivariate), suggesting that accounting for residual correlation between observations may not be essential if the time-course function has been appropriately modelled and parameters of interest are summary effects.

Conclusions: Time-course MBNMA provides a statistically robust framework for synthesising evidence on multiple treatments at multiple time-points. The methods can inform drug-development decisions, and provide the rigour needed in the reimbursement decision-making process, thereby acting as a bridge between early phase clinical research and HTA.

#### **4.10 Contributed - Methods & Theory: Nonlinear models**

**Tuesday 4 September 3.40pm - 4.40pm**

##### **A flexible sequential Monte Carlo algorithm for shape-constrained regression**

Kenyon Ng, Kevin Murray, Berwin Turlach

*The University of Western Australia*

Shape constraints, such as monotonicity or convexity, can be required on a regression curve to ensure its interpretability due to some external theory. Despite the importance of restricting the shape of the regression curves, existing shape constrained models are mainly concentrated on polynomial or spline models due to their relative ease of implementation. In this presentation, we propose an algorithm that is capable of imposing shape constraints on regression curves, without requiring the constraints to be written as closed-form expressions, nor assuming the functional form of the loss function. Our algorithm, which is based on Sequential Monte Carlo-Simulated Annealing, only relies on an indicator function that assesses whether or not the constraints are fulfilled, thus allowing us to enforce various complex constraints by specifying an appropriate indicator function without altering other parts of the algorithm. We demonstrate our algorithm by fitting rational function models subject to monotonicity and continuity constraints. Our results are comparable to that obtained from the constrained estimation particle swarm optimisation algorithm, which required an additional hyperparameter to operate.

#### **4.10 Contributed - Methods & Theory: Nonlinear models**

**Tuesday 4 September 3.40pm - 4.40pm**

##### **Beyond Beta regression: modelling bounded-domain variables in the presence of boundary observations**

Ioannis Kosmidis

*University of Warwick*

Beta regression is a useful tool for modelling bounded-domain continuous response variables, such as proportions, rates fractions and concentration indices. One important limitation of beta regression models is that they do not apply when at least one of the observed responses is on the boundary --- in such scenarios the likelihood function is simply 0 regardless of the value of the parameters. The relevant approaches in the literature focus on either the transformation of the observations by small constants so that the transformed responses end up in the support of the beta distribution, or the use of a discrete-continuous mixture of a beta distribution and point masses at either or both of the boundaries. The former approach suffers from the arbitrariness of choosing the additive adjustment. The latter approach gives a "special" interpretation to the boundary observations relative to the non-boundary ones, and requires the specification of an appropriate regression structure for the hurdle part of the overall model, generally leading to complicated models. In this talk we rethink of the problem and present an alternative model class that leverages the flexibility of the beta distribution, can naturally accommodate boundary observations and preserves the parsimony of beta regression, which is a limiting case. Likelihood-based estimation and inferential procedures for the new model are presented, and its usefulness is illustrated by applications.

#### **4.10 Contributed - Methods & Theory: Nonlinear models**

**Tuesday 4 September 3.40pm - 4.40pm**

##### **Non-Extensive Cross-Entropy Econometrics: introducing to the Nonlinear Simultaneous equation modelling**

Second Bwanakare

*University of Information Technology and Management in Rzeszow*

The aim of the paper is to present a new, Tsallis cross-entropy econometrics related estimator for a simultaneous nonlinear models. The presentation starts from the Kullback-Lebleir information divergence model, viewed as a connection between the Shannon entropy and the Bayesian philosophy. Targeting to go beyond the inconvenient ergodicity hypothesis imposed on the Linderberg's central-limit theorem-related applications, the next section proposes the power law(PL)-based non-extensive cross-entropy econometrics( NCEE). In the next step, recognizing the complexity of real social world where data generating system is generally represented by interconnected non ergodic phenomena, we propose a Seemingly Unrelated Nonlinear Regression model as a simultaneous equation case study. Concluding remarks end the presentation by indicating plausible new direction of research in this new scientific field.

## **4.2 Contributed - Official Statistics & Public Policy: UK Trade & Economy**

**Tuesday 4 September 3.40pm - 4.40pm**

### **Understanding asymmetries in UK bilateral trade data**

Marilyn Thomas, Adrian Chesson

ONS

The UK Trade project at the Office for National Statistics is an ambitious project which aims to strengthen the reliability and robustness of official UK trade statistics, as well as meet the emerging needs of users. A key focus of Phase 2 of the project is to identify, analyse and understand the causes of the trade data asymmetries which the UK holds with its key bilateral trading partners. In this regard, ONS is committed to disseminating information on asymmetries in both trade in goods and trade in services statistics. The outcome of this work will allow ONS to better understand the UK's bilateral trade data asymmetries in a global context, as well as communicate a more complete picture of the causes of asymmetries to users of trade data, and ultimately to improve official trade statistics.

## 4.2 Contributed - Official Statistics & Public Policy: UK Trade & Economy

Tuesday 4 September 3.40pm - 4.40pm

### Overseas Trade in Goods Statistics: Improving the way we process big, multidimensional datasets in ONS

George Zorinyants, Robert Breton, Andrew Banks  
*Office for National Statistics*

HMRC collects administrative data on trade in goods for every UK firm, covering the country of origin or destination, as well as the commodity exported or imported. Until recently, existing technology in ONS did not allow for data to be available at the lowest level of granularity possible. Instead, ONS was publishing aggregates: either data on a total product by country basis or a total country by product basis. We developed a new statistical process that preserves the full dimensionality of data, so that country by product trade in goods data, consistent with Balance of Payments Manual (BPM6), can be published. This will allow stakeholders to better understand the structure of trade with each partner country, and assess the potential impact of possible future changes of trade tariffs and regulations. As a second phase, we identify each firm according to its VAT registration number and link the trade data to an in-house register of businesses. This allows a further breakdown of trade by industry (based on a standard industrial classification), as well as an estimation of how many imports are used (e.g. as final consumption or investment). This could help users understand how trade affects the economy more widely. Possible future developments include further linking of the trade data to the ONS business register to reveal the structure of trade by region of the UK. Because of the size and complexity of data we used a distributed storage (HDFS, Hive) and computing (Spark) platform provided by Cloudera, Inc. We applied a number of data cleaning methods and developed a new methodology of disaggregation based on proximity and hierarchy of products. The new, more granular dataset is scheduled for publication in Blue Book 2018, and the first experimental statistics on trade industry breakdown are going to be released later this year.

### 4.3 Contributed - Applications of Statistics: Prediction

Tuesday 4 September 3.40pm - 4.40pm

#### Statistical modelling for assessing the risk of electricity shortfalls in Great Britain

Amy Wilson, Chris Dent, Stan Zachary  
*University of Edinburgh*

A reliable electricity supply is a key consideration for energy system planners. In Great Britain (GB), a planning study known as the electricity capacity assessment is performed annually to assess the long-term risk of insufficient electricity generation to meet demand. Around £700m per annum is spent on procuring electricity capacity as a result of this study. The need to meet climate targets has resulted in a rise in wind and solar generation in GB. As wind and solar generation are weather dependent, the effect of this rise has been an increase in variability in the availability of electricity generation. An additional difficulty is that wind, solar and demand are correlated. New statistical methods that account for this correlation and increased variability are therefore needed to assess the long-term risk of electricity shortfall. In this presentation, two new methods for modelling demand and wind when assessing the risk of an electricity shortfall in some future year under study will be described. The first method models the marginal distribution of demand-net-of-wind over the future year. The metric used to assess the risk is the expected number of hours of shortfall for the future year, which can be estimated from this marginal distribution. As data at times of shortfall are limited, extreme value theory is used to smooth the upper tail of demand-net-of-wind. The second method extends the first to account for correlations through time in demand-net-of-wind. This extension is necessary when there is electricity storage on the system, or when a wider range of metrics are required. A parametric model is used to capture the seasonal effects in demand-net-of-wind. A conditional dependence model for extremes is fitted to the residuals of this parametric model. Metrics can then be evaluated by simulating demand-net-of-wind for the future year using this model. The two methods will be compared, and results for the GB system presented.

### 4.3 Contributed - Applications of Statistics: Prediction

Tuesday 4 September 3.40pm - 4.40pm

#### Development of a model for predicting allergenicity of new and modified proteins

Tanja Krone,<sup>1</sup> Almar Snippe<sup>1</sup>, Lilia Babe<sup>2</sup>, Scott McClain<sup>3</sup>, Gregory Ladics<sup>2</sup>, Geert Houben<sup>1</sup>, Kitty Verhoeckx<sup>1</sup>

<sup>1</sup>TNO, <sup>2</sup>Dupont, <sup>3</sup>Syngenta

**INTRODUCTION**, The current allergy risk assessment of novel and modified food (proteins) relies mainly on the allergenicity risk assessment guidance for genetically modified plant foods. This guidance is based on a weight-of-evidence approach which assesses source of the gene, homology with known allergens, IgE binding, and stability in a pepsin resistant test. These tests mainly address cross reactivity and are not always suitable/available for novel proteins. Alternative testing strategies are therefore needed. This project aimed to develop a statistical model to predict if a protein is allergenic or not using protein specific characteristics.

**METHODS** Protein characteristics, such as % amino acids, instability index, possible S-S bridges, were obtained for half a million proteins from protein databases (SwissProt, UniProt) and mathematical programs (e.g. DIANNA, ProtParam). Proteins were assigned as allergenic when present in the allergen database (COMPARE). Random Forest analysis was used to find correlations between protein characteristics and the classification allergenic and non-allergenic, and to build a predictive model. Additional checks, such as a randomizations of outcomes and comparison of the important variables found using other statistical models, were performed to test the robustness of the model.

**RESULTS** 7 variables (Kingdom, GO-annotation membrane, stability index, secondary structure, percentage of Lysine, Arginine and Cysteine) were found relevant to predict if a protein is allergenic. Complications for model building lay in the choice and availability of protein characteristics data for allergenic as well as non-allergenic proteins. The final model was able to predict if a protein is an allergen with an accuracy of 88%, specificity of 89% and accuracy of 88. This result stays afloat after bias checks.

**CONCLUSION** A statistical predictive model based on protein characteristics was developed that is able to predict if a protein is allergenic or not with high accuracy. The variables used in the model are both statistically and biologically relevant.

### **4.3 Contributed - Applications of Statistics: Prediction**

**Tuesday 4 September 3.40pm - 4.40pm**

#### **Adoption and abandonment rules for economic evaluation of health care technologies: a Bayesian real options approach**

Daniele Bregantini,<sup>1</sup> Jacco Thijssen<sup>2</sup>

<sup>1</sup>*University of Liverpool*, <sup>2</sup>*University of York*

In this paper we propose a decision framework that allows for the assessment of the evidence about the cost-effectiveness of a healthcare technology when uncertainty and irreversibilities are present. The proposed model, while fitting the real option decision framework, allows for a Bayesian updating rule that makes the link between the evidence collected in clinical research and the expected payoffs of adoption to the health care system explicit. The model takes into account the cost of decision errors and explicitly models it in the payoff function. The implications in terms of opportunity costs of a decision taken with insufficient evidence are discussed and put into the real option context. Additionally, using a real-world cost-effectiveness study built on clinical trial evidence, we show how rules derived from our approach lead to quicker decisions when compared to the Value of Information decision framework, thus bringing higher payoffs and maximizing patient's health outcomes.

#### **4.4 Invited – RSS Prize Winners – Statistical Excellence for Early Career Writing 2018**

**Tuesday 4 September 3.40pm - 4.40pm**

##### **Preventing cancer: Mere rhetoric or a promising plan?**

Morten Valberg, Mats J. Stensrud

*Oslo Centre for Biostatistics and Epidemiology, University of Oslo and Oslo University Hospital*

There is an ongoing debate about the role of chance in cancer development, which was fuelled by Tomasetti and Vogelstein in *Science*: They claimed that the variation in cancer risk among tissues is mainly due to bad luck, not genes or environmental factors. However, we use twin data to show that the inequality in cancer risk across individuals is substantial: curiously, this variation is comparable to the variation in wealth across the World's population. Our findings suggest that pure randomness is not the most dominating cause of cancer, and that a large proportion of cancers could, at least in theory, be prevented.

## 4.5 Contributed - Methods & Theory: Design and testing

Tuesday 4 September 3.40pm - 4.40pm

### An outlier in an independent samples design

Ben Derrick

*University of the West of England, Bristol*

There is a flaw with some of the most commonly performed statistical tests. A paradox of the one sample t-test is the contrariwise decrease in the p-value as the value of an outlier increases in the direction of the overall effect [1]. Demonstration of this paradox is extended to the equal variances assumed and Welch's unrestricted to equal variances independent samples t-test. The phenomenon is explored using Monte-Carlo simulation, and compared with alternative two sample tests; the Mann-Whitney U test, and the Yuen-Welch t-test with 10% trimming per tail. Scenarios where the overall effect is concordant or discordant with the direction of the aberrant observation are considered. Sample data is generated under normality, with the subsequent inclusion of an aberrant observation in one sample. The aberrant observation is systematically varied. The total sample sizes for each of the two samples within a factorial design are {10, 15, 20}. The variances within the factorial design are {1, 4}. For each parameter combination, the proportion of 10,000 iterations where the null hypothesis is rejected is calculated at the 5% significance level, two sided. It is evidenced that the paradox for both forms of the independent samples t-test is exacerbated when the smaller sample size with the higher variance includes the aberrant observation, and as the imbalance between the sample sizes increases. Results also indicate that when the sample with the lower variance includes the aberrant observation, Welch's t-test and the Yuen-Welch t-test most closely retain Type I error robustness. Recommendations on choice of test for independent samples designs are given, ending with discussion on how these results impact analyses for partially overlapping samples designs. [1] Derrick, B., Broad, A., Toher, D., & White, P. (2017). The impact of an extreme observation in a paired samples design. *Metodološki Zvezki Advances in Methodology and Statistics*, 14(2), 1-17.

#### 4.5 Contributed - Methods & Theory: Design and testing

Tuesday 4 September 3.40pm - 4.40pm

##### **Sample Size Requirements for Validating a Reliable Risk Prediction Model using Binary Outcomes: A Simulation Study**

KHADIJEH Taiyari,<sup>1</sup> Gareth Ambler<sup>2</sup>, Rumana Omar<sup>2</sup>

<sup>1</sup>*MRC Biostatistics Unit*, <sup>2</sup>*Department of Statistical Science, University College London*

It has been suggested that when validating risk prediction models, there should be at least 100 events in data. However, few studies have examined the adequacy of this recommendation. There are also no guidelines regarding the sample size requirements when validating risk prediction models using hierarchical data, for example, when one has observations from several hospitals. The objective of this study is to conduct a simulation study based on real clinical data to investigate the sample size requirements for model validation in the context of both independent and clustered binary outcomes. A number of simulation scenarios were investigated by varying the number of events, the risk profile, the number of clusters, the cluster size and the intra-correlation coefficient (ICC). Standard or Random-intercept logistic regression models were fitted to the full-size development data and the quality of the resulting predictions were then quantified in validation datasets of varying size using measures of calibration, predictive accuracy and discrimination. For clustered outcomes, three types of predictions were considered; cluster-specific, population-averaged, and median predictions. From our study, the precision of predictive performance measures increases as the number of events in validation data increases. Moreover, the precision of predictive performance measures does not depend on the risk profile in the context of independent binary outcome. As well, the precision of those measures was affected differently by the number of events depending on the type of the predictions used. For example, the precision of calibration slope obtained using cluster-specific predictions was slightly better than that obtained using other prediction types. Furthermore, the calculated performance measures using cluster-specific predictions were biased at large ICC (say 20%). In conclusion, when validating a reliable risk prediction model using independent binary outcomes, one needs less than 100 events. However, one should have more events in validation data when validating a risk prediction model using clustered binary outcomes with large ICC.

## 4.5 Contributed - Methods & Theory: Design and testing

Tuesday 4 September 3.40pm - 4.40pm

### Location-adjusted Wald statistic for scalar parameters

Claudia Di Caterina,<sup>1</sup> Ioannis Kosmidis<sup>2</sup>

<sup>1</sup>*Department of Statistical Sciences, University of Padova*, <sup>2</sup>*Department of Statistics, University of Warwick, and the Alan Turing Institute*

Inference on a scalar parameter of interest is commonly constructed using a Wald statistic, on the grounds of the asymptotic validity of the standard normal approximation to its finite-sample distribution and computational convenience. A prominent example are the individual Wald tests for regression parameters that are reported by default in regression output in the majority of statistical computing environments. The normal approximation can, though, be inadequate, especially when the sample size is small or moderate relative to the number of parameters. In this work, the Wald statistic is viewed as an estimate of a transformation of the model parameters and is appropriately adjusted so that its null expectation is asymptotically closer to zero. The bias adjustment depends on the expected information matrix, the first-order term in the bias expansion of the maximum likelihood estimator, and the derivatives of the transformation, all of which are either readily available or easily obtainable in standard software for a wealth of well-used models. The finite-sample performance of the location-adjusted Wald statistic is examined analytically in simple models and via simulation in a series of more realistic modelling frameworks, including generalized linear models, meta-regression and beta regression. One application to brain data resulting from a neuroscience study is also considered. The location-adjusted Wald statistic is found able to deliver significant improvements in inferential performance over the standard Wald statistic, without sacrificing any of its computational simplicity.

## 4.6 Contributed - Communicating Statistics: Communicating Statistics in a Digital World

Tuesday 4 September 3.40pm - 4.40pm

### The RSS Champion Award for Excellence in Official Statistics Winner 2018 - The Future Farming and Environment Evidence Compendium

Jamie Jenkins, Jenny Kemp, Luke Ridley  
*Defra*

Informing Users Farmers do not always digest policy evidence, and many of the policy professionals working on the future offer were new to their roles, so a document was needed to educate them in order to ensure successful policy development. Several workshops were held to collaboratively develop a list of evidence questions to be included in the compendium. Communication of results Armed with these questions and mindful that many users spend little time browsing evidence, the compendium was developed into two layers – a visual summary of key messages and then detailed information that the user could reveal by clicking or touching that message. Feedback from the target audience demonstrated that the time taken to change language and explain difficult charts made it accessible. Colleagues in communications were pleased the design would translate across to social media platform. Problem Solving and Analysis One policy proposal was to put a limit on the amount of subsidy a single farm can receive, but sample limitations meant it was not possible to assess the impact of this using Defra's statistical data. To overcome this unmet need, discussions with the Rural Payments Agency resulted in access to the administrative data, allowing a full assessment of the number of farms impacted by limiting subsidy. Capturing Knowledge Input came from various analytical disciplines, including science, research, statistics and economics, to ensure a wide and robust evidence base. Information was gathered from many external organisations to give a balanced picture, included evidence from the Environment Agency, Forestry Commission, Natural England and Devolved Governments. Informing decisions The Defra analysts pushed the policy teams to link their policy proposals to the evidence base using page references to the compendium. The key proposals announced by the Secretary of State, Michael Gove, were built following analysis created and presented in the compendium.

## **4.6 Contributed - Communicating Statistics: Communicating Statistics in a Digital World**

**Tuesday 4 September 3.40pm - 4.40pm**

### **Meeting Census 2021 user needs: the development of an online flexible dissemination system**

Chris Ashford

*Office for National Statistics*

The Office for National Statistics is putting user needs at the heart of preparations for the dissemination of the 2021 Census. After a review of user feedback regarding their 2011 Census outputs experience, we have identified our strategic aims to improve the timeliness, accessibility and flexibility of the data delivery for the 2021 Census. An integrated ONS team from Census and Methodology are working in collaboration on an innovative and agile solution to develop an online flexible dissemination system for 2021 Census outputs. This system will use dynamic Statistical Disclosure Control methods to enable users to query and define their own 2021 Census outputs. Preliminary results include the development of a prototype system based on a 60 million record artificial database that contains personal and household variables at 7 different levels of geography. This presentation will outline how we aim to improve on the dissemination of 2011 Census outputs and successfully meet user needs for 2021, the dynamic Statistical Disclosure Control methodology used, its benefits and trade-offs and how we are involving users on the development of the flexible dissemination system.

## **4.6 Contributed - Communicating Statistics: Communicating Statistics in a Digital World**

**Tuesday 4 September 3.40pm - 4.40pm**

### **Transforming Health and Social Care Publications in Scotland**

David Caldwell, David Caldwell, Ewout Jaspers, Maighread Simpson  
*NHS National Services Scotland*

**Objectives:** Information Services Division (ISD) produces around 400 publications each year. Most of these are produced using SPSS, and published as static PDF documents with accompanying excel tables. Feedback has shown that our data can be challenging to find and digest in this format. Furthermore, production is time-consuming, often involving extensive manual formatting and checking. To improve user experience and make production more efficient and transparent, the transforming publications programme aims to modernise how ISD releases data.

**Methods:** We have been using a combination of Data Science, DevOps and user-driven design principles. We first identified our most common customer types and developed a set of personas. Customers were engaged with directly through interviews and focus groups to identify if our perceptions of customers' needs reflected reality, and the findings were translated into features for development. Once the team had an informed understanding of customer needs, we designed a new method of publishing data, focusing on one publication as a proof of concept. The team worked iteratively, involving customers to test and provide feedback on the new platform as it underwent development. To build this platform we utilised modern tools and techniques, including open-source software such as R, D3, git and GitHub.

**Results:** A prototype for releasing data was developed and released in December 2017. By co-designing a new model of presenting data, we can provide customers with the data they need in a way that they can understand. Furthermore, the new automated method of producing the publication has created time savings and reduced the risk of manual errors. We have encouraged continual feedback on the new publication, allowing the development of additional features which will help refine the product further. We are now working with a number of teams within ISD to transform their publications into this new design.

## **4.7 Contributed - Data Science**

**Tuesday 4 September 3.40pm - 4.40pm**

### **Data Science at Dŵr Cymru Welsh Water**

Kevin Parry

*Dwr Cymru Welsh Water*

The Data Science Team at Dŵr Cymru Welsh Water has grown substantially over the last couple of years, where it now seen as an integral component of the organisation. This talk will provide an overview of how the team has been able to support the organisation through the implementation of solutions and tools. The talk will provide some real examples of solutions that have been successfully deployed within the organisation, that are fundamentally based upon data science objects (such as machine learning models), and which our teams are using to improve compliance, manage risk and realise efficiencies. The talk will provide some information around where the team sits within the organisation and how it actively collaborates with business areas to identify where the implementation of certain data science approaches could bring substantial benefits. It will also cover how solutions are successfully deployed, including how complex and technical concepts, often associated with data science models, are translated in a manner that can be understood clearly by colleagues at all levels across the organisation. Some of the challenges that the team have encountered will be summarised, including how the team have successfully overcome these, often through close collaboration with stakeholders. The challenges around data quality and the conduciveness of data for analytics will be discussed, with some insight provided around how the organisation has embarked on a complementary Data Strategy, led by a dedicated Data Governance team, to improve the quality and usability of our data. The talk will conclude with an overview of how the team are looking to progress further in the near future (e.g. using new technologies, approaches) so to offer even further benefits to the organisation.

## 4.7 Contributed - Data Science

Tuesday 4 September 3.40pm - 4.40pm

### **Syndromic surveillance: the benefits of embedding a statistician within your team**

Roger Morbey, Alex Elliot, Andre Charlett, Gillian Smith  
*Public Health England*

Public Health England (PHE) is responsible for protecting the health of the nation against any potential threat from infectious disease or environmental exposure. Therefore, PHE monitors a wide range of data sources for signs of hazards. Prior to the London 2012 Olympics, the Health Protection Agency (HPA, PHE's predecessor) needed to improve its syndromic surveillance capabilities. Thus, creating a Games legacy of enhanced daily syndromic surveillance with automated statistical alarms. In order to improve statistical rigour, HPA created the post of "statistical project lead" within the multidisciplinary real-time syndromic surveillance team. Previously, the team had clinical and informatics expertise, but with statistical support from elsewhere in HPA providing ad hoc support. The novel approach of involving a statistician in all aspects of the team's work has had a number of advantages compared to relying on external experts or specialised software. For instance, an embedded statistician was able to gain a deep understanding of the complexities of syndromic data and their potential uses which may not be obvious to statistical consultants. Furthermore, the embedded statistician can get instant feedback on how well algorithms are working and proactively suggest new improvements. Similarly, whilst software packages can provide black box solutions for aberration detection, they have potential pitfalls if users cannot understand their working or adjust them when needed. PHE is now an international leader in syndromic surveillance, monitoring over 12,000 time series daily. We have created bespoke detection algorithms, and validated and improved them over several years. As a data science team, we have utilised informatics skills to automate data flows and epidemiological expertise to prioritise statistical alarms. We now use statistical tests alongside descriptive epidemiology in routine reporting and published research. Also, we are able to describe the complexities of real time syndromic surveillance data and the implications for statistical analysis to external collaborative researchers.

## **4.7 Contributed - Data Science**

**Tuesday 4 September 3.40pm - 4.40pm**

### **Demystifying big data in official statistics - it's not rocket science!**

Jens Mehrhoff

*European Commission*

The talk will initially define big data and discuss the interpretation in the area of official statistics. We will then focus on the use of big data in the production of official statistics, referring to the case study of electronic transactions data, better known as 'scanner data', for measuring the rate of change in consumer prices. As such, simple classification rules and similarity measures are introduced, which help in grouping items together. An empirical example shows how a price index can be calculated from this new data source. At all stages of the presentation two things are key: demystifying machine learning and the like, while, at the same time, highlighting the limits of what is technically possible. Looking beyond the production of official statistics, we will investigate the potential of big data for nowcasting and constructing new or complementary indicators. The final part will be devoted to quality issues, in particular coverage bias, and potential ways to deal with the situation.

## 4.8 Contributed - Industry & Finance: Banking culture and risk assessment

Tuesday 4 September 3.40pm - 4.40pm

### Why expected loss is not the right measure of credit risk: tail inference for balance sheet data

Andrei Sarychev

*European Central Bank*

The large bulk of risk assessment in the banking industry, including that for regulatory purposes, is performed using satellite models. These are founded on a number of implicit assumptions, of which the most tenuous one attributes all relevant variation in balance sheet outcomes to the contemporaneous variation in macroeconomic aggregates, and treats the unexplained variation as irrelevant noise. In the expected loss formulation, quantiles of the loss distribution are thus imputed from the quantiles of the state of the macroeconomy. This stands at variance with the anecdotal evidence on the financial crisis episodes, which purely macroeconomic shocks cannot explain. Empirically, this implicit assumption has not yet been rigorously tested. Indeed, it is impossible to test within the satellite model framework. To provide an alternative, I build a full-information probabilistic representation of the joint data-generating process for balance sheet outcomes and macroeconomic indicators. I specify a parsimonious state-space model and perform Bayesian inference on the marginal (as opposed to the conditional) distribution of the aggregate portfolio losses. I demonstrate a sizeable discrepancy between these distributions. Expressing the joint likelihood by means of a flexible vine-copula decomposition I show how the discrepancy is further amplified by the assumption of the Gaussianity of the statistical links between the macro and micro variables. These results suggest that credit portfolios may be substantially more risky than the current analytical practices would indicate. For regulators, the policy implication is that capital buffers based on macroeconomic stress tests are likely to be inappropriately low.

#### 4.8 Contributed - Industry & Finance: Banking culture and risk assessment

Tuesday 4 September 3.40pm - 4.40pm

##### ANALYSIS OF MICROFINANCE STRATEGIES FOR FINANCIAL RISK MANAGEMENT

Oyetayo Oluwatosin,<sup>1</sup> Olaniyi Mathew Olayiwola<sup>1</sup>, Abosede Seun Adeniran<sup>2</sup>

<sup>1</sup>*Federal University of Agriculture, Abeokuta, Nigeria*, <sup>2</sup>*Federal University of Agriculture, Ekiti State University, Ado-Ekiti*

Risk taking is described as an integral part of financial services. For micro-financing in particular, engaging in proactive risk taking is essential to their viability and long term sustainability. Maintaining a good strategy that ensures an optimal mix in risk-return trade-off is much more important for the microfinance banks (MFBs) that operate on a for-profit basis. Having faulted the value-at-risk technique which is common in the asset and liability literature, we introduce the multi-stage stochastic programming using econometric time series model. Specifically, for the scenario generation, we specify a VaR model with the inclusion of dichotomy regime which captures the multi-stage characteristics of assets. We use the liability derived investment (LDI) model to generate the liability series over the period of study. The optimization result shows that MFBs in Nigeria, are by far more risk averse than they are profit seeking. This comes with the attendant effect of not being able to achieve the outreach and sustainability objectives to the fullest. MFBs in Nigeria need to look into their investment strategy with a view to structuring the mix and value of the balance sheet components at different periods to meet their stated objectives

Key words: microfinance banks, asset and liability, risk, viability, sustainability

## 4.8 Contributed - Industry & Finance: Banking culture and risk assessment

Tuesday 4 September 3.40pm - 4.40pm

### What have we learnt assessing culture in UK banking?

Qamar Zaman, Michael Gardiner  
*Banking Standards Board*

*Co-presented with Michael Gardiner, Banking Standards Board*

The Banking Standards Board (BSB) exists to help raise standards of behaviour and competence across the UK banking sector, to the benefit of customers, clients, the economy and society as a whole. Over 2016 and 2017 the BSB conducted the largest ever assessment of behaviour, competence and culture in UK banking, involving 64,000 responses from those working within banks to a newly developed survey tool, with a further 1500+ people participating in detailed focus groups and interviews. This has allowed us to build a unique (and growing) data set. This talk will cover five aspects: How do we attempt to assess firms rigorously on a concept as 'fuzzy' as culture? The philosophical and empirical underpinning of the BSB model and tools How have we analysed the data gathered? From the use of ordinal logit, generalised ordinal logit, and mixed effects models, to employing grounded theory techniques in a systematic manner What do our results show? When there is a conflict between the values of a firm and the way it does business what explains it; why don't employees speak up; how the conception of what amounts to 'excessive' pressure is different in investment banking to other banking areas; where do women and men see things similarly and where do their perceptions differ; how do things change as life in banking progresses? How do we present information? A novel method for visualising logit results across multiple regressions with the aim of more easily observing and displaying patterns. What next? Extending and expanding the BSB Assessment; and, a next potential step of building a platform to conduct randomised control trials across banks.

## 4.9 Contributed - Environmental & Spatial Statistics

Tuesday 4 September 3.40pm - 4.40pm

### Detecting coherent changes in flood risk in the Great Britain

Ilaria Prosdocimi, Aoibheann Brady, Emiko Dupont  
*University of Bath*

Flooding is a natural hazard which has affected the UK throughout history, with significant costs for both the development and maintenance of flood protection schemes and for the recovery of the areas affected by flooding. The recent large repeated floods in Northern England and other parts of the country raise the question of whether the risk of flooding is changing, possibly as a result of climate change, so that different strategies would be needed for the effective management of flood risk. To assess whether any change in flood risk can be identified, one would typically investigate the presence of some changing patterns in peak flow records for each station across the country. Nevertheless, the coherent detection of any clear pattern in the data is hindered by the limited sample size of the peak flow records, which typically cover about 45 years. We investigate the use of multilevel hierarchical models to effectively use the information available at all stations in a unique model to better detect the presence of any sizeable change in the peak flow behaviour. Further we also investigate the possibility of attributing any detected change to naturally varying climatological variables. The advantages of using multilevel models over single at-site analysis and modelling conditions needed for these advantages to be realised will be discussed.

## 4.9 Contributed - Environmental & Spatial Statistics

Tuesday 4 September 3.40pm - 4.40pm

### Statistical Analysis of Long Term Seasonal Rainfall Variations in a Southern Caribbean Island

Aruna Rajballie<sup>1</sup>, Kiran Tota-Maharaj<sup>2</sup>, Amarnath Chinchamee<sup>3</sup>, Vrijesh Tripathi<sup>1</sup>

<sup>1</sup>*The University of the West Indies, St. Augustine, Trinidad and Tobago*, <sup>2</sup>*University of the West of England (UWE Bristol), Bristol, United Kingdom*, <sup>3</sup>*The University of Trinidad and Tobago, Port of Spain, Trinidad and Tobago*

Rainfall is a renewable water resource and precipitation patterns are critical to providing sustainable water supplies for several Caribbean Small Island Developing States (SIDS). The Republic of Trinidad and Tobago (T&T) is a twin island republic southernmost country in the Caribbean region and has vast amounts of freshwater resources. However, these sources are subject to depletion based on recent inconsistent precipitation. With climate change, it is projected that annual rainfall over the islands may decrease, temperatures and evaporation rates can increase and hence water resources from ponds, rivers, dams and streams decrease as well as reductions in aquifer recharge. Fluctuations in annual rainfall have therefore caused a change in the timing and lengths of the wet and dry seasons and an average annual reduction in water levels across reservoirs and dams. This variation in rainfall activity can impact the country's ability to meet growing water demands of the population. An understanding of rainfall variability across different regions in T&T is essential for strategic planning regarding future sustainable water demand. This project examines the variations in rainfall across both islands. Time series analysis will be used to model seasonal, monthly and annual rainfall data across T&T for a 15 year period from 2000-2014. Spatial analysis of data will be used to determine whether there is any association between rainfall and supply of water in these Southern Caribbean regions. The models are expected to be used to forecast future rainfall patterns and values which may be beneficial to Government institutions, the local water resources agency and other public utilities regarding upgrades and development of new infrastructure such as reservoirs or Desalination plants and groundwater aquifers all of which can improve the supply of water to its customers.

## 4.9 Contributed - Environmental & Spatial Statistics

Tuesday 4 September 3.40pm - 4.40pm

### Statistical analysis of weather-related property insurance claims

Christian Rohrbeck,<sup>1</sup> Jonathan Tawn<sup>1</sup>, Deborah Costain<sup>1</sup>, Arnaldo Frigessi<sup>2</sup>  
<sup>1</sup>Lancaster University, <sup>2</sup>University of Oslo

We consider the association between the number of property insurance claims and a set of weather metrics, including precipitation and snow-melt levels. Weather events which cause severe damages are of general interest; decision makers want to take efficient actions against them while the insurance companies want to set adequate premiums. The modelling is challenging since the underlying dynamics are highly complex, due to potential threshold and interaction effects, and vary geographically due to differences in topology, construction designs and climate. Rohrbeck et al. (2018) (*Annals of Applied Statistics* 12, pages 246-282) introduce a statistical framework which comprises both mixture and extremal mixture modelling, the latter being based on a discretised generalised Pareto distribution. Moreover, the authors introduce a temporal clustering algorithm which aggregates claims over consecutive days based on observed weather metrics, and derives more informative explanatory variables. Their results for three Norwegian cities show that the combination of the statistical model and the clustering algorithm leads to an improved model fit and captures the spatial dependence between locations. However, the complexity of the approach may produce highly uncertain model estimates when applied to rural regions. We address this limitation in this talk. First, we summarise the existing approach. Then, we introduce a spatial clustering algorithm which aggregates claims across regions if these are likely to be related to the same weather event. By combining our spatial cluster algorithm with the existing methodology, we achieve a better understanding of the association between claims and weather events. The methodology is applied to multiple regions across Norway to illustrate its benefits.

## **Keynote 3 – Significance Lecture**

**Tuesday 4 September 5.10pm – 6.10pm**

### **Hannah Fry's Guide to Being Human in the Age of Algorithms**

Hannah Fry

*University College London*

For decades, human activities and decisions have been supported by algorithms. They are the hidden rules and instructions that help our computers to process data and run complex calculations. But in recent years, algorithms have moved from a supporting to a starring role. As our machines have become more powerful, the algorithms have become more sophisticated – so much so that they are now in control of potentially life-changing decisions. In the courts, algorithms decide if jail time is warranted. In hospitals, they match organ donors to waiting patients. And on the streets, they steer driverless cars. In each of these scenarios, wrong decisions can lead to tragic outcomes.

Ahead of the publication of her new book, *Hello World*, Dr Hannah Fry explores our relationships with algorithms, the responsibilities we give them, and the impact they are having on our societies – including the good, the bad, and the downright ugly.

## 5.1 Medical: Recent developments in growth trajectory charts for infants

Wednesday 5 September 8.30am - 9.50am

### Assessing infant weight gain with growth charts – thrive lines and correlation surfaces

Tim Cole

*UCL Great Ormond Street Institute of Child Health*

Weight centile growth charts are widely used in paediatrics for diagnosis and management. They allow serial weight measurements to be plotted so that the weight centile and also the gain in weight, the change in centile over time or “centile crossing”, are visualised. The centiles are straightforward to interpret, but centile crossing is not – the chart cannot show if a given rate of centile crossing is common or uncommon. Some years ago I developed “thrive lines” to help interpret centile crossing, using a dataset with measurements every 4 weeks from birth to 52 weeks. They are based on the concept of conditional weight gain, the regression of current weight on previous weight, expressed as z-scores  $z_2$  and  $z_1$  respectively. To mark slow weight gain, the 5th centile for current weight conditional on weight 4 weeks earlier is predicted to be  $E(z_2) = r.z_1 + 1.64 \sqrt{(1-r^2)}$  where  $r$  is the correlation between  $z_1$  and  $z_2$ . The thrive lines are a set of lines added to the chart that join up successive weights 4 weeks apart as linked by the formula, and the slopes of the lines at each age indicate the rate of downward centile crossing for the slowest-growing 5% of infants. However, using a 4-week period is restrictive, and being able to use more general time intervals would usefully extend the thrive line concept. This needs the correlations between pairs of weights to be known for quite general measurement ages. The talk will describe a method to use longitudinal weight data to construct such correlations as a smooth surface, a 3D representation of the correlation between weights measured at any pair of ages within a given age range. The underlying algebra, an interesting mix of correlation and regression, will be developed, and practical examples will show correlation surfaces for weight in infancy and height in puberty.

## **5.1 Medical: Recent developments in growth trajectory charts for infants**

**Wednesday 5 September 8.30am - 9.50am**

### **Post-natal weight gain in preterm-born children- a method assessing conditional trajectory**

John Lowe, Sarah Kotecha, John Watkins, Sailesh Kotecha  
*Cardiff University*

**Objectives:** Rapid postnatal weight gain is associated with adverse respiratory outcomes in childhood. However, the preterm-born population (gestation of <37 weeks' at birth) is less well studied perhaps because of the challenges of defining adequate growth using standard growth charts. We assessed if increased respiratory symptoms are associated with rapid weight gain in infancy in this vulnerable group.

**Study design:** We used data from our cohort of preterm- and term-born (n=3,425) children, aged 1-10 years. Respiratory outcomes obtained from a parent-reported respiratory questionnaire were regressed on conditional postnatal weight velocities calculated from birth until 9 months of age, then adjusted for covariates.

**Results:** Rapid infant weight gain was associated with increased wheeze-ever (OR1.30 95% CI 1.14, 1.49), recent wheeze (OR1.21 95% CI 1.05, 1.41) and inhaler usage (OR1.24 95% CI 1.08,1.36). However, only wheeze-ever was robust to the addition of covariates to the model (OR1.17 95% CI 1.01, 1.37).

**Conclusions:** This study suggests that postnatal growth rates are important for future respiratory health in preterm-born children. Since they are unlikely to have the same etiology of lung disease as their term-born peers, it is important to consider strategies for ensuring appropriate growth. Optimising nutrition in the neonatal period and beyond presents a potential, but challenging, intervention.

## 5.1 Medical: Recent developments in growth trajectory charts for infants

Wednesday 5 September 8.30am - 9.50am

### **Growing Preterm Infants in the United States Proportionally ... the Past, the Present, and the Future**

A. Nicole Ferguson

*Kennesaw State University*

Our team has previously published contemporary intrauterine curves with body mass index (BMI, Olsen, 2015) as a measure of body proportionality. We found BMI proved more suitable for the modern U.S. population than Lubchenco's ponderal index curves. In addition, we showed BMI is theoretically the best measure of proportionality (Ferguson, 2018). Recently our team created longitudinal median BMI preterm growth curves as a complement to our cross-sectional curves (Williamson, 2018) using a large US sample (n=68,693). We stratified infants by gender, gestational age (GA) at birth (24-27, 28-31 and 32-36 weeks GA), and quintiles of birth BMI. Both GA at birth and postnatal age influenced changes in BMI. The first 7 to 14 days following birth had an initial drop followed by a near linear increase in BMI for all GA groups and both genders. A comparison within GA but across the five percentile groups revealed that the top percentile group showed the greatest weight loss during nadir. After nadir, the curves for all percentile groups within GA and both genders followed a similar growth pattern and remained consistently below the cross-sectional curves. For both genders, the extremely preterm infants (GA 24-27 weeks) had a more rapid increase in BMI compared to the cross-sectional curves. This group returned to the optimal growth level over their stay in the NICU. Because BMI in infancy can affect health throughout an individual's life, it is essential to understand whether NICU practices are successful in helping infants reach and maintain a healthy BMI. We are preparing research showing small, appropriate, or large for gestational age by BMI (SGA, AGA, and LGA, respectively) do not agree with weight-based classifications. Furthermore, the level of agreement between the two methods differs across gestational ages at birth, raising questions about the relationship between duration in the NICU and BMI.

## 5.1 Medical: Recent developments in growth trajectory charts for infants

Wednesday 5 September 8.30am - 9.50am

### On reconsidering early life growth trajectories for Premature children

William Watkins<sup>1</sup>, Mallinath Chakraborty<sup>2</sup>, Daniel Farewell<sup>3</sup>, Sujoy Banerjee<sup>4</sup>

<sup>1</sup>*Department of Infection and Immunity, Cardiff University,* <sup>2</sup>*Regional Neonatal Intensive Care Unit, University Hospital of Wales, Cardiff, UK,* <sup>3</sup>*Department of Population Medicine, Cardiff University,* <sup>4</sup>*Neonatal Intensive Care Unit, Singleton Hospital, Swansea, UK*

**Introduction:** Current UK reference charts for monitoring post-natal growth of preterm infants were created using data at birth with the LMS method. As post-natal growth is significantly different from in-utero growth this may lead to the wrong diagnosis of growth failure in most preterm infants, with the added risk of inappropriate nutritional interventions and potential long-term adverse effects. Using longitudinal data, we aimed to create predictive models of post-natal weight-gain for preterm infants.

**Methods:** Daily weight data of preterm infants born between 23-31 weeks' gestation from birth to discharge (or death) at 14 Welsh neonatal units between 2011 and 2014 was obtained from the neonatal electronic database. Approximately 2/3rd of cleaned data was used to teach the model and the other 1/3rd to test it. We developed a mixed model where infant weight was regressed on gender, time since birth and time since conception. Time since birth was included as a spline. The model was hierarchical (repeated measurements nested within children) to account for correlation within a child's weight measurements. The model variance was recalculated periodically to correctly represent the increasing variation in weight over time. The model was run separately for 3 gestation bands – 23-25 weeks, 26-28 weeks and 29-31 weeks allowing growth charts of percentiles to be determined for any gestation within each band and either gender.

**Results:** The test data was better fitted by the new percentiles than those from the LMS at all gestations between 23 and 31 weeks.

**Conclusion:** We believe that this approach represents the first step towards a flexible methodology to appropriately model post-natal growth for premature infants, and should be extendable by taking additional morbidities into account. This should allow updateable personalised growth charts for preterm infants to more accurately assess a preterm infant's post-natal progress.

### **5.3 Applications of Statistics: Statistics and Genetics: A fruitful relationship**

**Wednesday 5 September 8.30am - 9.50am**

#### **Inferring causality from genome wide association studies**

Toby Johnson

*GlaxoSmithKline*

Over the last decade, genome wide association studies (GWAS) have revolutionized our knowledge of DNA sequence variants that have strong and robust statistical associations with human diseases and traits. This is primarily because of increased sample sizes, which in turn are due to lower costs of genotyping and DNA sequencing, broad adoption of collaborative meta-analytic approaches, and widespread access to data from the UK Biobank. In larger sample sizes, genetically related individuals will be sampled almost surely, and making full use of these data motivates finding computationally tractable ways to fit billions of random effects models. Although genetic associations are typically not susceptible to reverse causation or uncorrected confounding, it remains a challenging and unduly neglected problem, to make inference about which genes or proteins (as opposed to DNA sequence variants) are causal for human diseases. Some fully or approximately Bayesian methods show promise in this regard. I will describe our approaches to solving these problems, with application to selecting and validating targets for future drug discovery projects.

### **5.3 Applications of Statistics: Statistics and Genetics: A fruitful relationship**

**Wednesday 5 September 8.30am - 9.50am**

#### **Statistics and Genetics: A Fruitful Relationship - A brief history of the relationship between the two disciplines**

Heather Cordell

*Newcastle University*

In this presentation, I shall briefly summarize the history of the relationship between the two disciplines of Statistics and Genetics, with a focus on particular selected highlights. Starting from the work of Gregor Mendel in 1866 (rediscovered independently at the turn of the century by Correns, de Vries, von Tschermak-Seysenegg and Spillman), it soon became clear that inherited traits obey simple statistical rules. However, the simple rules of Mendelian inheritance seemed at odds with the biometric approaches developed by Francis Galton (and promoted by Raphael Weldon and Karl Pearson), until R.A. Fisher's seminal 1918 paper (and subsequent work arising) effectively resolved the conundrum. Some specific areas in which genetics has impacted statistics, or, conversely, statistics has impacted genetics, will be discussed. These include epistasis, likelihood-based techniques (including REML), sequential testing, Approximate Bayesian Computation (ABC) and False Discovery Rate (FDR).

### **5.3 Applications of Statistics: Statistics and Genetics: A fruitful relationship**

**Wednesday 5 September 8.30am - 9.50am**

#### **Statistical challenges of the "Let's Measure Everything" era of human genetics**

Luke Jostins-Dean

*University of Oxford*

Over the last decade, human geneticists have discovered a very large number of genetic variants that impact "headline" human traits (e.g. risk of common diseases). In parallel to this, our ability to measure broader "genome-adjacent" human phenotypes has exploded: researchers often measure expression levels of genes across multiple tissues, epigenetic marks, metabolites, gut microbe abundances, etc. Geneticists analyse these diverse "-omics" datasets in order to draw causal pathways from genetic variants, through cellular and systemic phenotypes, through to headline traits. This new "Let's Measure Everything" approach to human genetics raises a new set of statistical challenges. In this talk, I will discuss some examples of these challenges, and the attempts that have been made to address them. I will discuss the "large p, small N" problem in analysing gene expression data, where a large number of genes are tested using a small number of samples, and how variance stabilisation techniques have allowed these experiments to proceed successfully. I will also discuss the problem of modelling covariance between large numbers of observed variables in the presence of potential confounding, including cases where no single unique solution exists, and how penalised likelihood approaches can provide solutions. Finally, I will discuss some currently unsolved problems in statistical genetics, and how new developments in statistics could help drive genetics forward in the future.

## 5.4 Social Statistics: Bayesian applications in social contexts: dealing with data difficulties

Wednesday 5 September 8.30am - 9.50am

### Bayesian Methods in Demography Using Child Labour as an Exemplar

Wendy Olsen

*University of Manchester Department of Social Statistics*

Our main aim is to estimate the prevalence of child labour (defined as a damaging situation of overwork, hazardous work, or forced labour, when faced by children of school age) in South Asian countries. Under age 16, all child labour except very light work is illegal in international law based on UN and ILO conventions. Child labour is always under-reported, for three reasons: exaggeration of school attendance; hiding of child labour by some employers and parents; and country level regulations that define 15-year-old children as eligible to work fulltime. All three sources of undercount can help frame our Bayesian exploration of the prevalence and causes of child labour in each region of six South Asian countries: Bangladesh, Bhutan, India, Myanmar, Nepal, and Pakistan. We combine a range of random-sample household data sources from International Labour Organisation (ILO) and the United Nations (Multiple Indicator Cluster Survey) to model child labour. These are recent cross-sectional data. Each child's age is a polynomial covariate in a Bayesian Poisson model of sample-survey counts (Woodward, 2012: *Bayesian Analysis Made Simple: An Excel GUI for Winbugs*). We control for associated factors including location (rural/urban), child's gender, gender of the head of household, minority religious or ethnic group, and region within the country. Three advantages of Bayesian estimation are demonstrated: 1) Weak prior assumptions about undercount can be used informatively. 2) More efficient estimation for each region is achieved by pooling all the regions. 3) Asymmetric high density intervals can be graphed, creating a shaded density of conditional child labour risk over age of child. Nepal and parts of Pakistan have very high levels of child labour using these estimation methods. The results are considerably altered, depending on the definition of 'child work': A) employment in the paid labour market gives low estimates, while (B) considering more informal activities gives much higher estimates.

## **5.4 Social Statistics: Bayesian applications in social contexts: dealing with data difficulties**

**Wednesday 5 September 8.30am - 9.50am**

### **Bayesian Analysis of Social Network Data from Patchy Sources**

Johan Koskinen

*University of Manchester*

Social network analysis is typically concerned with the study of social ties (links) between individuals (nodes) and the patterns of interactions that emerges out of the collection of ties. The canonical form of network data is the so-called 'network census', often obtained from the roster method, whereby the presence or absence of ties is elicited for all pairs of nodes in a fixed, and predetermined set of nodes. By necessity, however, many researchers have to rely on various sampling approaches for collecting their network data. These sampling approaches – ego-centric network data collection and various forms of link-tracing designs – provide data different from the canonical 'network census'. In this talk we briefly outline some of the challenges associated with trying to analyse the structure of networks using data collected in these unstructured ways. We present a Bayesian data-augmentation approach for analysing sampled of incomplete network data that asserts a specific log-linear model for the network ties that is called an exponential random graph model. The approach is illustrated by way of a number of empirical examples, based on current active projects in a range of different disciplines.

## **5.5 Methods & Theory: Variable selection in high-dimensional and non-standard settings**

**Wednesday 5 September 8.30am - 9.50am**

### **Exponential Family Principal Component Analysis**

Luke Smallman,<sup>1</sup> Andreas Artemiou<sup>1</sup>, Jennifer Morgan<sup>2</sup>, Paul Harper<sup>1</sup>, William Underwood<sup>2</sup>  
<sup>1</sup>*Cardiff University*, <sup>2</sup>*NHS Wales Delivery Unit*, <sup>3</sup>*Oxford University*

In this talk we will explore a number of methods for extending principal components analysis to exponential-family data, with text data (modelled using a Poisson distribution) as a motivating example. In particular, we will look at how two existing methods can be sparsified for improved performance, and discuss a unifying framework for all such methods.

## **5.5 Methods & Theory: Variable selection in high-dimensional and non-standard settings**

**Wednesday 5 September 8.30am - 9.50am**

### **Post Selection Inference for Stepwise Regression Using Multiple Comparisons**

Kory Johnson

*University of Vienna*

Forward stepwise regression provides an approximation to the sparse feature selection problem and is used when the number of features is too large to manually search model space. In this setting, we desire a rule for stopping stepwise regression using hypothesis tests while controlling a notion of false rejections. That being said, forward stepwise regression is commonly considered to be "data dredging" and not statistically sound. As the hypotheses tested by forward stepwise are determined by looking at the data, the resulting classical hypothesis tests are not valid. We present a simple solution which leverages classical multiple comparison methods in order to test the stepwise hypotheses using the max-t test proposal of \cite{BujaB14}. The resulting procedures are fast enough to be used in high-dimensional settings. Other procedures estimate new, computationally difficult p-values and have significant lower power. Furthermore, our proofs readily extend to more general correlation learning methods such as Sure Independent Screening.

## **5.5 Methods & Theory: Variable selection in high-dimensional and non-standard settings**

**Wednesday 5 September 8.30am - 9.50am**

### **On the Model Selection Properties of the Lasso**

Ulrike Schneider

*Vienna University of Technology*

We present an explicit formula for the correspondence between the Lasso and the least-squares estimator in low dimensions and derive analogous results for the relationship between the Lasso estimator and the quantity  $X'y$  in high dimensions without any assumptions on the regressor matrix. Given these results, we also investigate the model selection properties of the Lasso estimator based on geometric conditions and show that possibly only a subset of models might be selected, completely independently of the response vector. Finally, we present a condition for uniqueness of the estimator in this context that is necessary as well as sufficient.

## 5.8 Industry & Finance: Design and analysis of industrial and network experimentations

Wednesday 5 September 8.30am - 9.50am

### Closed-loop automatic experimentation for optimisation

Timothy Waite<sup>1</sup>, Dave Woods<sup>2</sup>

<sup>1</sup>*University of Manchester*, <sup>2</sup>*University of Southampton*

Automated experimental systems, involving minimal human intervention, are becoming more popular and common, providing economical and fast data collection. We discuss some statistical issues around the design of experiments and data modelling for such systems. Our application is to “closed-loop” optimisation of chemical processes, where automation of reaction synthesis, chemical analysis and statistical design and modelling increases lab efficiency and allows 24/7 use of equipment. Our approach uses nonparametric regression modelling, specifically Gaussian process regression, to allow flexible and robust modelling of potentially complex relationships between reaction conditions and measured responses. A Bayesian approach is adopted to uncertainty quantification, facilitated through computationally efficient Sequential Monte Carlo algorithms for the approximation of the posterior predictive distribution. We propose a new criterion, Expected Gain in Utility (EGU), for optimisation of a noisy response via fully-sequential design of experiments, and we compare the performance of EGU to extensions of the Expected Improvement criterion, which is popular for optimisation of deterministic functions. We also show how the modelling and design can be adapted to identify, and then down-weight, potentially outlying observations to obtain a more robust analysis.

## **5.8 Industry & Finance: Design and analysis of industrial and network experimentations**

**Wednesday 5 September 8.30am - 9.50am**

### **A graph-theoretic framework for algorithmic design of experiments**

Ben Parker,<sup>1</sup> Vasiliki Koutra<sup>2</sup>, Steven Gilmour<sup>2</sup>

<sup>1</sup>*University of Southampton*, <sup>2</sup>*King's College London*

How can we design experiments on networks well, and how can we use networks to design experiments? In this paper, we demonstrate that considering experiments in a graph-theoretic manner allows us to exploit automorphisms of the graph to reduce the number of evaluations of candidate designs for those experiments, and thus find optimal designs faster. We show that the use of automorphisms for reducing the number of evaluations required of an optimality criterion function is effective on designs where experimental units have a network structure. Moreover, we show that we can take block designs with no apparent network structure, such as one-way blocked experiments, row-column experiments, and crossover designs, and add block nodes to induce a network structure. Considering automorphisms can thus reduce the amount of time it takes to find optimal designs for a wide class of experiments.

## **5.8 Industry & Finance: Design and analysis of industrial and network experimentations**

**Wednesday 5 September 8.30am - 9.50am**

### **Designing Experiments with Unstructured Treatments for General Network Structures**

Frederick Kin Hing Phoa,<sup>1</sup> Ming-Chung Chang<sup>2</sup>, Jing-Wen Huang<sup>3</sup>

<sup>1</sup>*Academia Sinica*, <sup>2</sup>*National Central University*, <sup>3</sup>*National Tsing Hua University*

Experiments on connected units are commonly conducted in various fields, such as agriculture trials, medical experiments and social networks. In these applications, an experimental unit connects to one another, and the treatment applied to a unit has an effect, called a network effect, on the responses of the neighboring units. Designing such experiments was rarely discussed in the literature. Parker, Gilmour, and Schormans (2017) initiated a study of As-optimal designs on connected experimental units with unstructured treatments, assuming that the network effects were unknown constants. In this work, we studied a similar design problem under an assumption that the network effects were random effects. It led to a property that the responses of two units were correlated if some neighbors of one unit and those of the other received the same treatment. Alphabetical optimality criteria were considered for selecting good designs with high efficiency of estimating the treatment effects and/or high accuracy of predicting the network effects. We provided theoretical conditions for designs to be optimal and illustrate our theory with some numerical examples.

## **5.9 Environmental & Spatial Statistics: Environmental monitoring**

**Wednesday 5 September 8.30am - 9.50am**

**Monitoring marine contaminants: part statistics, part data management, part communication.**

Rob Fryer

*Marine Scotland*

The Oslo and Paris Commission coordinates the assessment of contaminant levels and trends in the North Sea and north-east Atlantic. It is a big assessment: last year, we analysed over 24000 time series submitted by 11 countries over a 35 year period, and synthesised the results for 14 biogeographic regions. I'll describe the methods that we have developed to routinely analyse so many time series, and the pragmatic decisions required along the way. Challenges include the usual issues of data quality, incorporating 'historic' data submitted when there were different reporting requirements, dealing with evolving (resource driven) monitoring objectives and strategies, and incorporating the effects of improved chemical methods with lower limits of detection. I'll also show how we present the results online using an application that provides a regional overview accessible to managers and stakeholders and that allows the user to drill down to the raw data behind each time series.

## 5.9 Environmental & Spatial Statistics: Environmental monitoring

Wednesday 5 September 8.30am - 9.50am

### Now you see them, now you don't: the importance of modelling butterfly populations

Emily Dennis

*Butterfly Conservation*

Butterfly populations are undergoing various changes which require investigation. They respond sensitively and rapidly to changes in habitat and climate, hence their population status is a valuable indicator for changes in biodiversity. Butterflies are the most comprehensively monitored invertebrate taxa. Count data from the UK Butterfly Monitoring Scheme is used to derive abundance indices, which form one of the UK Government's 18 biodiversity indicators. Opportunistic citizen science records are used to describe butterfly distributions. Devising suitable statistical methods for modelling butterfly abundance presents challenges. Butterflies have multi-stage life cycles and hence count data fluctuate within each year in response to their emergence as adults. Many species are also multivoltine, with up to three broods of adults emerging each year. Furthermore, fitting models to data for many species from extensive, long-term monitoring schemes can be challenging and computer intensive. We present recent models that describe seasonal variation in count data of the adult stage, and demonstrate their application to a number of species. In particular a generalised abundance index approach (Dennis et al, 2016; *Biometrics*, 74, 1305-1314) is very efficient due to the use of concentrated likelihood techniques and provides new parametric descriptions of seasonal variation which produce estimates of parameters relating to emergence and survival. Dynamic models (Dennis et al, 2016; *JABES*, 21, 1-21) explicitly describe dependence between broods and years to produce indices and estimated productivities separately for each brood. We provide examples of new developments and modifications, for example to migrant butterfly species and to two difficult to distinguish Skipper species.

## **5.9 Environmental & Spatial Statistics: Environmental monitoring**

**Wednesday 5 September 8.30am - 9.50am**

### **People and birds: Analytical challenges from ecological citizen science data**

Alison Johnston

*Cornell University and Cambridge University*

Citizen science data are increasingly being used in ecological research. However, these data often contain a number of challenges: 1) spatially biased locations; 2) biases towards certain species; 3) considerable variation in effort and other survey characteristics; 4) considerable variation in observer expertise. In order to draw robust ecological conclusions, it is important that analysts consider all of these challenges that are inherent to many citizen science datasets. There are two main solutions to these challenges: imposing a post-hoc structured protocol onto the dataset for design-based inference or including covariates in a model for model-based inference. We outline some statistical and analytical approaches to these challenges.

## **RF1: Rapid Fire Talks**

**Wednesday 5 September 10am - 10.50am**

### **Mapping the urban forest**

Philip Stubbings

*ONS Data Science Campus*

In collaboration with the ONS Natural Capital Accounts team, the Data Science Campus have developed an automated method for estimating the amount of publicly accessible trees and vegetation across 112 major towns and cities in England and Wales. An end-to-end image processing pipeline has been developed which can automatically detect vegetation utilising recent advancements in deep learning from images periodically sampled from Google street view. This rapid fire talk aims to cover some of the main techniques utilised and challenges encountered during the development of this project.

## RF1: Rapid Fire Talks

Wednesday 5 September 10am - 10.50am

### **Detecting signals for adverse drug reactions (ADRs) using non- constant hazard in longitudinal data: a comparison of three model-based methods**

Victoria Cornelius,<sup>1</sup> Odile Sauzet<sup>2</sup>

<sup>1</sup>Imperial College London, <sup>2</sup>Bielefeld University

Objective: Pharmacovigilance monitors the drug harm profile to identify new ADRs or changes in frequency of known ADRs. Historically data from spontaneous reporting systems have been the mainstay but the last decade has seen increased interest in use of electronic health records (EHRs) for pharmacovigilance. Recently proposed signal detection methods, that make optimal use of the additional information in EHR, still rely on an 'expected' value to be calculated which required a large database of drugs. The causal mechanism of an ADR will often result in the occurrence being time dependant. We propose flagging signals based on detecting a variation in hazard of an event using a time-to-event approach. This has added advantage that no control group required. Methods: Cornelius et.al. 2012 method based on the Weibull Shape Parameter (WSP) has been demonstrated to work well if ADRs occur at the beginning or end of an observation period. We describe and compare two additional models. Performance is demonstrated on an EHR Bisphosphonates dataset and through a simulation study (population 2,500–10,000, adverse event (AE) and ADR rates 1-10% and 0.01% -10% respectively). Results: We describe the development of a test which uses a mixture exponential and Weibull model (mWSP), and a second test which is based on a power generalised Weibull distribution (pWSP) introduced by Bagdonavicius and Nikulin. For both models we use numerical maximisation of the likelihood which, under certain assumptions, can provide confidence intervals on which to base a test for constant hazard. In the Bisphosphonates example, the pWSP and WSP test correctly signalled two ADRs, and correctly did not signal two adverse events not associated with the drug. The mWSP did not identify any signals and in simulation had a high proportion of non-convergence. Conclusions: The power generalised Weibull (pWSP) offers a practical alternative to Weibull shape parameter test (WSP).

## **RF1: Rapid Fire Talks**

**Wednesday 5 September 10am - 10.50am**

### **Predictive Modelling: What's the customer's propensity to pay?**

Vasiliki Bampi

*Dwr Cymru Welsh Water*

Welsh Water continuously strives to provide the best level of service to its customers. The Propensity to Pay project supports this objective by helping to identify customers that are struggling to pay the bills, to provide them with appropriate support. For the purposes of this project, a predictive model was developed that identifies the likelihood of a customer having difficulty paying their bill. Variable reduction techniques were used to identify the most significant predictors and undersampling methods were applied to boost the performance of the model. A logistic regression model was implemented to identify the most important attributes, and performance measures such as Gini coefficient and ROC curve were used to evaluate the model. The output of the model is a risk score between 0-100. Customers with high scores tend to be more likely to be in debt in compare to those with low scores. The complexity and size of the data required by the model provided some initial challenges that were successfully overcome. Data manipulation methods were undertaken using SQL and R was used for statistical analysis and model development. The output of this model could help us understand better our customers' needs. By identifying risky customers we can act proactively to assist customers with paying their bills.

## **RF1: Rapid Fire Talks**

**Wednesday 5 September 10am - 10.50am**

### **A New Gamma- Generalized Gamma Mixture Cure Fraction Model: Application to Ovarian Cancer at The University College Hospital, Ibadan, Nigeria**

Serifat Folorunso

*Department of Statistics, University of Ibadan*

In some therapeutic studies, attention is needed on the number of patients who are not susceptible to the event of interest (recurrence of disease) and expected to be cured. In this article, the emphasis is on estimation of the proportion of patients who are cured based on mixture cure model usually used to model failure time data with long-term survivors. Parameter estimates under several most commonly used parametric models, namely lognormal, loglogistic, Weibull and generalized Gamma distributions were explored. The gamma link function was a generator that formed the proposed gamma generalized gamma mixture cure model; which incorporates the model that can handle heavily skewed survival data. The maximum likelihood estimation (MLE) approach is employed to estimate the model parameters. Ovarian cancer data and a simulation study were used for assessing the efficiency and capability of the proposed over the existing one. Our results show that the cure fraction estimates from the proposed model is found to be quite robust.

## RF1: Rapid Fire Talks

Wednesday 5 September 10am - 10.50am

### On the sample mean after a group sequential trial

Ben Berckmoes<sup>1</sup>, Anna Ivanova<sup>2</sup>, Geert Molenberghs<sup>3</sup>

<sup>1</sup>*Universiteit Antwerpen*, <sup>2</sup>*KU Leuven*, <sup>3</sup>*KU Leuven and UHasselt*

A popular setting in medical statistics is a group sequential trial with independent and identically distributed normal outcomes, in which interim analyses of the sum of the outcomes are performed. Based on a prescribed stopping rule, one decides after each interim analysis whether the trial is stopped or continued. Consequently, the actual length of the study is a random variable. It is reported in the literature that the interim analyses may cause bias if one uses the ordinary sample mean to estimate the location parameter. For a generic stopping rule, which contains many classical stopping rules as a special case, explicit formulas for the expected length of the trial, the bias, and the mean squared error (MSE) are provided. It is deduced that, for a fixed number of interim analyses, the bias and the MSE converge to zero if the first interim analysis is performed not too early. In addition, optimal rates for this convergence are provided. Furthermore, under a regularity condition, asymptotic normality in total variation distance for the sample mean is established. A conclusion for naive confidence intervals based on the sample mean is derived. It is also shown how the developed theory naturally fits in the broader framework of likelihood theory in a group sequential trial setting. A simulation study underpins the theoretical findings.

## **RF1: Rapid Fire Talks**

**Wednesday 5 September 10am - 10.50am**

### **Using Spatio-temporal clustering to model antibiotic use**

Jonathan Ansell

*The University of Edinburgh Business School*

Analysing data over both space and time is an issue in various areas of application including health, marketing and public services. Based on the spatio-temporal clustering algorithm ST-DBSCAN, we describe a method of clustering spatial time series while taking into account varying data point densities across space in a continuous manner via density-based distance weighting. The resulting clusters can not only inform decision-making through a deeper understanding of spatio-temporal data, but also be used for representative sampling of data and the generation of synthetic data sets. Our method is developed using data from National Health Service (NHS) Scotland Open Data on drug prescription data. Possible applications reach further, e.g. for retailers and public services striving for an increased understanding of their customers while, at the same time, being concerned about retaining anonymity of identifiable single-person data. Our results demonstrate how, and offer a solution for, the necessity of methods adaptive to varying densities when performing spatio-temporal clustering of data points over large spatial areas. Further research is planned to develop an approach that allows for changes in the size of considered spatial areas ('zooming'), as well as for changes in cluster composition and memberships over time.

## RF2: Rapid Fire Talks

Wednesday 5 September 10am - 10.50am

### **Development and validation of a clinical prediction model to identify adult patients (aged 18 – 50) with type 1 diabetes requiring early insulin therapy**

Anita Grubb,<sup>1</sup> Angus Jones<sup>2</sup>, Kashyap Patel<sup>2</sup>, Beverley Shields<sup>2</sup>, Richard Oram<sup>2</sup>, Katharine Owen<sup>3</sup>, Andrew Hattersley<sup>2</sup>

<sup>1</sup>University of Exeter Medical School, <sup>2</sup>National Institute for Health Research Exeter Clinical Research Facility, <sup>3</sup>Oxford Centre for Diabetes Endocrinology and Metabolism, University of Oxford, Churchill Hospital, Oxford

**Background and aims:** Correctly determining diabetes subtype at diagnosis is important to ensure optimal treatment and education, but is often difficult, particularly in young adults, where misclassification is common. We aimed to develop a clinical prediction model combining clinical features and GAD autoantibodies (marker of type 1 immune process) to accurately identify patients with type 1 diabetes (T1D), requiring early insulin therapy.

**Methods:** We studied 1,352 participants in Exeter-based cross-sectional cohorts diagnosed with diabetes when aged 18-50 years. Our study outcome was T1D which was robustly defined using blood C-peptide and rapid insulin requirement (< 3 years). We developed two prediction models using logistic regression based on: 1) clinical features (age at diagnosis and BMI); 2) clinical features and GAD. Discrimination and calibration performance of the models were estimated using internal bootstrap validation. External validation of the models was performed using 701 participants taken from the Young Diabetes in Oxford (YDX) study.

**Results:** T1D was present in 13% of participants in the Exeter cohorts. The model combining clinical features was highly discriminative (c-statistic 0.90 [95% CI 0.88, 0.93]); internal bootstrap validation showed a small optimism (0.0006). Adding GAD improved discrimination (c-statistic 0.96 [0.95, 0.97]) with optimism 0.0009. Hosmer-Lemeshow test for calibration was non-significant in both models (p=0.95 & 0.39). T1D was present in 19% of participants in the YDX study. In the external validation, both models still showed excellent discrimination (clinical features c-statistic 0.86 [0.82, 0.89]; clinical features + GAD c-statistic 0.92 [0.89, 0.95]). However, there was evidence of miscalibration in the high risk deciles (p = 0.004 & 0.007).

**Conclusion:** This is the first study to show that a clinical prediction model combining clinical features and GAD can accurately identify T1D in a group of patients where misclassification is most common. Our model has excellent discrimination and routine use of this model in clinical practice is likely to reduce misclassification.

## **RF2: Rapid Fire Talks**

**Wednesday 5 September 10am - 10.50am**

### **What do newly developed ONS business surveys tell us about the UK economy?**

Chloe Gibbs, Michael Hardie  
*Office for National Statistics*

ONS is working towards providing more granular data on the UK economy, in response to recent reviews of economic statistics. To meet these review recommendations, and to improve the quality of Supply-Use tables and National Accounts, the ONS reinstated the Annual Purchases Survey (APS) in 2015, and developed the Annual Survey of Goods and Services (ASGS), which launched in 2017. The introduction of both surveys has provided the ability to analyse the supply and use of individual UK businesses, at a level of product detail not previously possible. Prior to the new surveys, limited product detail was available, with several assumptions used to breakdown industry totals. We will undertake microdata analysis of both sources to investigate the supply and use of UK businesses. This is important to reflect the UK economy's changing composition, as well as understanding UK business characteristics and the interdependencies between them. Linking these new data sources to well-established surveys, such as the Annual Business Survey (ABS) and UK Manufacturers' Sales by Product Survey (Prodcom), will be considered, to provide a full picture of the UK economy. The analysis will focus on several areas, including the predominant products produced outside a business's main industrial classification, and the diversity of different industries. The intermediate inputs of key industries, such as Manufacturing, will also be investigated and, for industries and businesses appearing in both APS and ASGS, the inputs required to produce certain outputs. The APS and ASGS also provide a measure of imports and exports, respectively, with ASGS collecting a detailed product breakdown of the services exported. An investigation of the differing composition of output in the domestic economy compared to the export economy, will be particularly important as the UK prepares to exit the European Union. This analysis is ongoing but will be completed in advance of the conference, where the results and conclusions will be shared.

## **RF2: Rapid Fire Talks**

**Wednesday 5 September 10am - 10.50am**

### **Sparsity and network measures: an application towards a new index measuring the extent of Common Ownership in a given industry.**

Nicoletta Rosati

*Joint Research Center - European Commission*

The existence of common shareholders among competing companies in an industry is raising the concern of academics and policy makers, given that its possible effects on market efficiency and competition have been investigated only in a small number of specific industries, and no consensus has been reached about a more general effect on the economy. The most popular tool used to assess the effects of common ownership (CO) is the Modified Herfindahl-Hirschman Index, which allows to measure the distortion introduced in market competition due to the presence of common owners. Nevertheless, it fails to measure directly the extent of CO itself. In general, the empirical studies have limited the measurement of CO to mainly descriptive measures of specific characteristics of the market structure, without presenting a unified index determining the overall extent of the phenomenon. We consider some possible methodologies that may be applied to construct an index of the extent CO, coming from the literature on sparsity and on networks, respectively. The concept of sparsity is generally found in the literature on inequality or diversity, and studies the distribution of a certain phenomenon in a given population. This framework can be applied to the CO analysis, looking at the distribution of the investments of shareholders across firms in a given market. Social networks study the empirical structure of the relationships between elements of a population. The corporate ownership structure of a market can be represented through a network, where the relations are given by the ownership links between firms and shareholders. A series of measures taken from both approaches are applied to the study of CO, and illustrated through a simple example. Measures that allow comparison with benchmark scenarios are also considered. Sparsity measures and other matrix norms are also applied to the network representation. The methodologies will be tested later on real data.

## **RF2: Rapid Fire Talks**

**Wednesday 5 September 10am - 10.50am**

### **Back-calibration from splines**

David Wille

GSK

Smoothing splines are widely used in many areas of applied analysis and statistics. Whereas predictors – functions for returning the splines value for a given independent value – are widely implemented, methods for their inverse, reading back independent values for a given value of the spline are more problematic. The implementation of many packages is often complex as it is not always immediately obvious where the appropriate parameters can be found. This talk provides one simple solution to this problem. Noting that splines are between knots locally ordinary polynomials, and that their parameters can be uniquely estimated by sampling appropriate predicted values, we show how our problem can be resolved using standard dedicated polynomial root-finding software without recourse to more general convergence-dependent methods. Our talk will be illustrated by a collection of functions written in R and will conclude on the comparison of two methods, a linear model or the solution of a Vandermonde system, for the estimation of the required parameters.

## **RF2: Rapid Fire Talks**

**Wednesday 5 September 10am - 10.50am**

### **Implicitly weighted robust classification for high-dimensional data**

Jan Kalina

*Institute of Information Theory and Automation of the Czech Academy of Sciences*

Standard data mining procedures are sensitive to the presence of outlying measurements in the data. Therefore, robust data mining procedures are highly desirable, which are resistant to outliers. In this work, we propose new robust classification procedures for high-dimensional data with the number of variables exceeding (perhaps largely) the number of observations. At the same time, algorithms for the efficient computation of the novel methods are proposed. Particularly, we use the idea of implicit weights assigned to individual observation to propose several robust regularized versions of linear discriminant analysis, suitable for data with the number of variables exceeding the number of observations. The approach is based on a regularized version of the minimum weighted covariance determinant estimator and represents a unique attempt to combine regularization and high robustness, allowing to down-weight outlying observations. Classification performance of new methods is illustrated on real data sets from various applications.

## RF2: Rapid Fire Talks

Wednesday 5 September 10am - 10.50am

### Fisher's Fundamental Theorem: it's not just about genetics

Nicholas Galwey  
*GlaxoSmithKline*

R.A. Fisher's Fundamental Theorem of Natural Selection states that 'The rate of increase in fitness of any [species of] organism at any time is equal to its genetic variance in fitness at that time.' Its meaning has been the subject of debate, because it depends on the definition of the terms 'fitness' and 'genetic variance'. It is closely related to the 'growth-rate theorem' which states that 'In a subdivided population the rate of change in the overall growth-rate is proportional to the variance in growth rates.' This presentation will express this theorem in a form that connects naturally to a graphical representation of the probability distribution of fitness, and will show that, when 'fitness' is appropriately defined, its variance and rate of change are equal, not merely proportional, as follows. Let  $f(W)_t$  be the probability density of fitness  $W$  at time  $t$ , such that after a time interval  $\Delta t$ ,  $f(W)_{t+\Delta t} = Wf(W)_t/l$ , where  $l = \int_0^\infty Wf(W)_t dW$  is the integral of  $Wf(W)_t$  from  $W = 0$  to infinity. It is also necessary to specify that  $E(W)_t = 1$ . The presentation will argue that in this formulation, 'fitness' equals the sum of additive genetic effects on reproductive success over all loci in an individual's genome. The other components of the individual's reproductive success – effects of dominance, epistasis, environment and genotype  $\times$  environment interaction and stochastic effects – comprise departures from this additive component, and their expected value in the next generation is zero. Fisher's theorem is then seen to describe a particular example of a phenomenon common to other contexts – notably economics – where the diversity among competing entities is expected to decrease over time, so that the rate of increase in the mean competitive ability becomes progressively smaller.

### **RF3: Rapid Fire Talks**

**Wednesday 5 September 10am - 10.50am**

**The design and analysis of sequential trials with binary outcomes, when errors in achieving and measuring the stimulus are present.**

Mike Hicks, Sam Ellis, Kevin Stone  
*Defence Ordnance Safety Group*

There are numerous designs that generate the suggested stimuli for sequential trials. Such suggestions may be based on initial estimates, and/or the outcomes of previous stimuli using simple heuristics or optimal design criteria. Analysis methods for the resulting data range from the maximisation of the likelihood for binary data to the methods that are available when formulated as a generalised linear model. We investigate the performance of different combinations of design and analysis method when used to find the stimulus level that corresponds to a probability of success/failure e.g. the velocity of an armour piercing projectile that gives a 95% chance of penetrating a given thickness of armour plate or the dose required in an in-vivo experiment required to give a 50% probability of a pre-defined end point within a fixed time period. For each we are interested in:

- How the errors in the stimulus, either due to the methodology or to the practicalities of achieving the stimulus, affect the accuracy of the parameter estimates.
- How robust they are to initial estimates, including assumptions about the underlying population distribution.

This has involved the simulation of trials for multiple sample sizes for different design methods including the simulation of deterministic and/or probabilistic errors in the suggestions and their achievement. And the analysis of trials data using various estimation techniques including comparisons of the estimates arising from the different optimality criteria. We give a brief history of the design and analysis methods and an overview of their benefits and drawbacks. Some detail regarding the efficiencies of separating the data generation and the data analysis procedures within the simulation. And provide a summary of the simulation results comparing the various design methods against a known answer. The aim is to provide a definitive guide to the benefits and drawbacks of each of the methods with a view to identifying the most reliable approaches.

## RF3: Rapid Fire Talks

Wednesday 5 September 10am - 10.50am

### Adherence to the MIND Diet predicts a lower risk of Mild Cognitive Impairment in the Framingham Heart Study

Jayandra Himali<sup>1</sup>, Alexa Beiser<sup>2</sup>, Debora Melo-van-Lent<sup>3</sup>, Ramachandran Vasani<sup>4</sup>, Paul Jacques<sup>5</sup>, Sudha Seshadri<sup>6</sup>, Matthew Pase<sup>7</sup>

<sup>1</sup>Boston University, <sup>2</sup>Boston University School of Public Health, Boston, MA, USA, <sup>3</sup>German Center for Neurodegenerative Diseases DZNE, Bonn, Germany, <sup>4</sup>Boston University School of Medicine, Boston, MA, USA, <sup>5</sup>Tufts University, Boston, MA, USA, <sup>6</sup>Glenn Biggs Institute for Alzheimer's & Neurodegenerative Diseases, UT Health San Antonio, TX, USA, <sup>7</sup>Swinburne University of Technology, Australia

Background: Adherence to the Mediterranean-DASH diet Intervention for Neurodegeneration Delay (MIND) diet has shown promise as a strategy to promote healthy cognitive aging. However, such findings require replication in independent cohorts. Accordingly, we examined the prospective association of adherence to the MIND diet with the risk of incident mild cognitive impairment (MCI) in the Framingham Study.

Methods: We studied 1,424 dementia-, stroke-, and MCI-free participants (mean age 69[SD,5], 47% men) followed on average for 10[SD,5] years. Dietary intake was assessed at three time points using a validated semi-quantitative food frequency questionnaire. Dietary components were scored in accordance with the original MIND diet papers. Of 15 dietary components 10 are considered healthy and five are unhealthy. Overall scores ranged from 0-15. We generated a MIND diet score by averaging across 3 examination cycles (exams 5[1991-1995], 6[1995-1998], and 7[1998-2001]). Surveillance for incident MCI commenced at exam 7 and continued for up to 17 years, during which we observed 178 MCI events. Diagnosis of MCI was by Petersen's criteria. We used a series of proportional hazards models to relate the cumulative MIND diet score to the risk of incident MCI. Mind diet scores were examined as a continuous variable and by tertiles.

Results: After adjustment for age, sex, and calorie intake, each unit increase in the MIND diet score was associated with a decreased risk of MCI (HR: 0.88[95% CI: 0.81,0.97]). Persons with MIND diet scores in the top (0.58[0.40,0.85]) but not middle tertile (0.79[0.56,1.12]) also had a lower risk of MCI, as compared to the bottom tertile. Results were comparable following adjustments for physical activity and vascular risk factors. We did not observe any effect modification by ApoE4 allele status.

Conclusions: Higher adherence to the MIND diet was associated with lower risk of MCI. Further studies are needed to elucidate whether adopting the MIND diet can alter cognitive trajectories with aging.

## **RF3: Rapid Fire Talks**

**Wednesday 5 September 10am - 10.50am**

### **Sequence balance minimisation: new minimisation procedure for clinical trials using unequal treatment allocation ratios**

Vichithranie Madurasinghe

*Queen Mary University London*

Background: Minimisation is a widely used randomisation scheme that ensures excellent balance between treatment groups for several prognostic factors, even in small samples. However, extending it to trials using unequal allocation ratios is challenging. Sequence balance minimisation is a new minimisation procedure for trials using unequal allocation ratios seeking to achieve balance across several prognostic factors.

Methods: Treatment and factor balancing properties of sequence balance minimisation was explored in a simulation study for a two arm trial with 1:2 allocation ratio. Number of prognostic factors on which to achieve balance ranged from 1 to 10 prognostic factors with 2 levels occurring in equal probabilities. Sample size was set to 30, 60 and 120. For comparison, treatment and factor balance achieved using stratified block randomization (a long established randomization scheme used for achieving better balanced treatment groups in clinical trials) was also examined.

Results: Number of prognostic factors (upto 10) included in the sequence balance minimisation scheme had little impact on overall treatment and factor balance achieved. The mean and median number of subjects allocated to each treatment group was as same as the number expected, with variability in allocations achieved increasing slightly as the number of prognostic factors increased, when the probabilities assumed for random element decreased, and also when the sample size increased. The Mean and median factor imbalance remained tightly around zero even when the chosen factor was not included in the minimisation scheme, though the variability was greater. The variability in factor imbalance increased slightly as the random element decreased, as the number of prognostic factors and sample size increased. With stratified block randomisation, as the number of prognostic factors increased treatment and factor balance deteriorated substantially.

Conclusions: Results show that sequence balance minimisation has good treatment and factor balancing capabilities, and particularly useful for small trials seeking to achieve balance across several prognostic factors.

## **RF3: Rapid Fire Talks**

**Wednesday 5 September 10am - 10.50am**

### **Mallows Model Averaging Estimator for Support Vector Regression**

Francis Kiwon, Jeffrey Racine  
*McMaster University*

Support Vector Machines (SVM) are widely applied in a regression setting for the prediction of values of the real continuous response variables (Drucker et al. 1996; Smola and Scholkopf 2004). There is a recent heightened interest in the use of model selection and averaging methods that can further reduce the mean squared error of regression, and model selection has recently been extended to Support Vector Regression (SVR). These model selection procedures utilize the generalized cross-validation criterion (Gunter and Zhu 2007), or tailor the principles of minimizing leave-one-out error bounds for SVM classifiers (Chang and Lin 2005). For model averaging, Bayesian model averaging has been used on a set of candidate models involving a regularization path algorithm and Occam's Window method (Wang and Liao 2012). As a frequentist alternative to existing Bayesian model averaging techniques, we propose the extension of a Mallows model averaging (MMA) procedure, which weighs each of candidate models by minimizing a Mallows criterion (Hansen 2008). First, given the data-generating process we order models suggested by cross-validation over a set of tuning parameter values and kernel families, from lowest to highest generalization error. A quadratic program is then used to obtain the solution vector. Monte Carlo replications reveal that in a range of settings where we modify the signal-to-noise ratio, sample size, and the form of the data-generating process, the MMA estimator outperforms that based on model selection. We also consider the proposed approaches' performance on some real-world data sets.

## **RF3: Rapid Fire Talks**

**Wednesday 5 September 10am - 10.50am**

### **Managing Discolouration Risk in a Water Distribution Network through Machine Learning and Data Visualisation**

Isaac Bowen

*Dwr Cymru Welsh Water*

One of Dŵr Cymru Welsh Water's key responsibilities is to ensure that the water provided to our customers is of an acceptable colour as well as ensuring it is safe to drink. The cause of discoloured water is usually caused by the disturbance of material deposits within a pipe network by some form of hydraulic event. These events may be planned or unexpected, and may include essential pipe maintenance or third party usage. We distribute over 800 million litres of water every day to our customers through 27,000 km of water pipes, meaning the process of managing the risk of customers receiving discoloured water can be very complex. It is unfeasible to remove material deposits from the whole network, so targeted interventions in the most at risk areas ensures we achieve the maximum positive impact. This project aimed to produce a model to help understand and forecast discolouration risk throughout our network to aid proactive preventative operational activity. A number of data sources were used to develop the model including flow, water quality samples, operational activity, network structure and discolouration contact history spanning several years. Collating these data sources allowed the successful development and testing of a random forest machine learning model to predict the expected number of contacts in the upcoming week within small operating areas. A final solution was then developed by combining model predictions with the key risk indicators identified through feature importance. This dashboard displays temporal and spatial visualisations to help network operators make informed decisions about where to perform interventions to minimise the impact on customers. Our tool is currently undergoing a trial period to gather feedback and help estimate its impact on live environment before a company-wide rollout.

### **RF3: Rapid Fire Talks**

**Wednesday 5 September 10am - 10.50am**

#### **Developing a predictive risk model for anxiety after treatment for breast cancer: applying best practice in prognostic statistics to inform psycho-oncology**

Jenny Harris,<sup>1</sup> Edward Purssell<sup>2</sup>, Jo Armes<sup>3</sup>, Emma Ream<sup>3</sup>, Victoria Cornelius<sup>4</sup>

<sup>1</sup>King's College London / University of Surrey, <sup>2</sup>King's College London, <sup>3</sup>University of Surrey, <sup>4</sup>Imperial College London

**Objective:** Primary care services play an important role in managing the supportive care of breast cancer (BC) survivors. Those at risk of delayed or recurrent psychological effects are not easily identified and are often discharged by specialist teams if assessed as low risk of cancer recurrence. Prediction risk models (PRMs) are useful in identifying those at increased risk but are seldom used in practice. Current approaches to PRM development and validation in psycho-oncology are poor. We present our methodological framework for developing and testing of PRMs using an example in psycho-oncology.

**Methods:** Data from the Supportive Care Needs Survey of >800 women with BC at the end of treatment and 6 months later was used to demonstrate good-practice framework for model development. The outcome was the hospital anxiety and depression scale- anxiety (HADS-A). Missing data were handled using multiple imputation using chained equations (MICE). LASSO and stepwise variable selection were compared to a full linear model using MI and complete-case (number, precision and variance explained). Model performance was assessed and bootstrapping used to obtain estimates for internal validation.

**Results:** Twenty candidate predictors were considered, missingness ranged from <1% to 20%, with 18% for HAD-A. MICE with predictive mean matching for continuous variables and 50 imputations was undertaken. The full linear MI regression model included all 20 variables (R<sup>2</sup> 0.60) and MI LASSO identified 5 (R<sup>2</sup> 0.60). LASSO and Stepwise complete-case identified 9 variables (R<sup>2</sup>0.58) and 7 variables (R<sup>2</sup>0.59). MI LASSO was preferred for its parsimony and was supported by bootstrap estimates.

**Conclusions:** Most of the statistical techniques described are widely available but underutilised in psycho-oncology. Model validation and implementation studies are now required. This PRM framework may prove useful to inform routine nursing practice and supportive care for breast cancer patients.

## RF4: Rapid Fire Talks

Wednesday 5 September 10am - 10.50am

### **Identification of Disease Subtypes Using Multivariate Longitudinal Data: a Comparison of Growth Curve Mixture Models and a Two-Stage Cluster Analysis Approach**

Benjamin Leiby,<sup>1</sup> Md Jobayer Hossain<sup>2</sup>

<sup>1</sup>Thomas Jefferson University, <sup>2</sup>Nemours Biomedical Research, A.I. DuPont Children's Hospital

In many chronic conditions, the vast majority of patients will experience stability in their disease characteristics, and models that consider the trajectory of symptoms in a sample representative of the population as a whole may identify little to no change over time. The presence of smaller subgroups who have experienced symptomatology that is substantially different from the majority is often suspected or known anecdotally from clinical experience. Identification of these subtypes and early predictors of progression are important for better classification of disease and potential development of earlier interventions. This presentation is motivated by a modestly-sized (n=161) cohort of patients with glaucoma who were followed annually over a 4-year period. Multiple clinical, functional, and quality of life measures were collected each year. As is common in studies of ophthalmology, most clinical measurements are taken on both eyes. Growth curve mixture models are a commonly-used model-based approach to identifying subgroups of patients based on longitudinal data. We apply univariate and multivariate growth curve mixture models to identify subsets of patients with differing outcome trajectories. We discuss the challenges of applying multivariate growth curve mixture models to data of this type, including relaxing typical assumptions of conditional independence to account for potential residual correlation among the same outcome measured in both eyes and equality of residual variances across latent classes. We also discuss the challenges posed by moderate sample size when the subtype is relatively rare. We consider a recent alternative that applies conventional cluster analysis methods to empirical best linear unbiased predictors from linear mixed effects models. Through data analysis and simulation, we identify the advantages and disadvantages of both approaches.

## RF4: Rapid Fire Talks

Wednesday 5 September 10am - 10.50am

### On using predictive-ability tests in the selection of time-series prediction models

Robert Kunst<sup>1</sup>, Mauro Costantini<sup>2</sup>

<sup>1</sup>*Institute for Advanced Studies*, <sup>2</sup>*Brunel University*

Comparative ex-ante prediction experiments over expanding subsamples are a popular tool for the task of selecting the best forecasting model class in finite samples of practical relevance. Flanking such a horse race by predictive-accuracy tests, such as the test by Diebold and Mariano (1995), tends to increase support for the simpler structure. We are concerned with the question whether such simplicity boosting actually benefits predictive accuracy in finite samples. We consider two variants of the DM test, one with naive normal critical values and one with bootstrapped critical values, the predictive-ability test by Giacomini and White (2006), which continues to be valid in nested problems, the F test by Clark and McCracken (2001), and also model selection via the AIC as a benchmark strategy. Our Monte Carlo simulations focus on basic univariate time-series specifications, such as linear (ARMA) and nonlinear (SETAR) generating processes. In this work, we strongly emphasize the design of convenient visual summaries of the various simulation results. The bottom line of the study is that the additional testing stage tends to make the forecast model selection too conservative, in the sense that the heightened support for the simpler structure fails to be optimal. We also point out some cases where the simplicity boosters help in the model selection decision. In short, AIC proves to be a strong benchmark, whereas the Giacomini-White test cannot be recommended for this type of model selection. The other test procedures that involve bootstrap perform well, but the additional complexity and processing time would not support their usage over a simple training-sample comparison.

## RF4: Rapid Fire Talks

Wednesday 5 September 10am - 10.50am

### **Determining the extent of web-based intervention use in health research: a systematic review**

Elena Koneska, Duncan Appelbe, Susanna Dodd  
*Institute of Translational Medicine, University of Liverpool*

Background: Web-based interventions in clinical trials are on the increase as they can provide accessible support to patients without the need to travel to a clinic and have enormous potential to improve health and healthcare delivery. [1] High quality usage data are required for accurately evaluating the impact of public health interventions and to inform efficacy of the intervention. [2] However, the process of defining and measuring intervention “dose” is not necessarily straightforward or obvious, and the accuracy of methods used to assess web usage is unknown. This project examines the extent of use of online interventions in randomised controlled trials as a first step to determine accurate ways to measure online intervention use.

Methods: We are conducting a systematic search of the literature in order to identify all published reports of randomised controlled trials that have involved an online intervention. A piloted data extraction form is being used to record information on whether web usage was recorded as part of these studies, and if so, the methods used to gather web usage data and whether these data were used to inform efficacy analyses.

Results and conclusions: Results from the systematic review will be presented, demonstrating the extent of use of web-based interventions in randomised trials to date and the methods used to record, and adjust for, web usage information in health research. This systematic review will be used as a basis for further investigations relating to the issue of how to record web usage in health research.

[1] Marina B. Wasilewski, Jennifer N. Stinson, Jill I. Cameron, *International Journal of Medical Informatics*, Volume 103, Issue null, Pages 109-138 , PMID: 28550996 [2] Hong Chen, David Hailey, Ning Wang, Ping Yu, *Int J Environ Res Public Health*. 2014 May; 11(5): 5170–5207. Published online 2014 May 14. doi: 10.3390/ijerph110505170, PMCID: PMC4053886

## RF4: Rapid Fire Talks

Wednesday 5 September 10am - 10.50am

### **Graphical methods to aid assessment of item performance in monitoring developmental outcomes of children aged 0-3 years in low and middle income countries**

Gillian Lancaster<sup>1</sup>, Gareth McCray<sup>1</sup>, Melissa Gladstone<sup>2</sup>, Patricia Kariger<sup>3</sup>, Vanessa Cavallera<sup>4</sup>, Tarun Dua<sup>4</sup>, Magdalena Janus<sup>5</sup>

<sup>1</sup>Keele University, <sup>2</sup>University of Liverpool, <sup>3</sup>UCLA Berkeley, USA, <sup>4</sup>World Health Organisation, <sup>5</sup>McMaster University, Canada

The aim of the fourth United Nations Sustainable Development Goal (SDG) is to promote equity in access to early education and requires the monitoring of learning and developmental outcomes in young children. Reliable methods of measurement are therefore crucial to monitor achievement of the SDG. Currently, there is a dearth of low-cost, cross-culturally validated, simple to implement tools that can be used in low and middle income settings for the most sensitive and youngest age range, 0-3 years. The indicators of Infant and Young Child Development (IYCD) tool was developed by a WHO-led team to address this gap. The IYCD is a 100 item-tool, divided into 5 age ranges, within 4 domains: motor, language-cognitive, socio-emotional, and general behaviour. It is a caregiver interview that can be administered with the help of a tablet, providing audiovisual aids especially useful in low-literacy settings. We used detailed statistical analyses of data from field test studies in Brazil, Malawi, and Pakistan to determine item performance, reliability and validity of the tool as a caregiver report (in contrast to a direct assessment), and item response theory to devise a Development for Age (DAZ) score. The results were presented to child development experts using graphical methods as an aid to reaching consensus on the final selection of items. The analyses revealed excellent age-specific developmental distributions across settings, thus confirming the cultural neutrality of the tool. Association with socio-demographic and family factors further supported its validity. While most children under 3 in low and middle income countries are not in preschools, their earliest development vastly contributes to their learning trajectories and outcomes in later life. Developmental assessment tools are essential for monitoring progress at the population level.

## **RF4: Rapid Fire Talks**

**Wednesday 5 September 10am - 10.50am**

### **Evidence evaluation and functional data analysis**

Ya-Ting Chang, Colin Aitken<sup>1</sup>, Grzegorz Zadora<sup>2</sup>, Patrick Buzzini<sup>3</sup>, Geneviève Massonnet<sup>4</sup>  
<sup>1</sup>*University of Edinburgh*, <sup>2</sup>*Institute of Forensic Research*, <sup>3</sup>*Sam Houston State University*,  
<sup>4</sup>*University of Lausanne*

A mathematical formulation is developed for a likelihood ratio as a measure of support for a proposition over another when the data are functions. The propositions considered are those of common source and of different sources for two sets of data where one of them has a known source of origin and the other does not. Hierarchical models that take account of between- and within-source variation are developed. B-splines are used in the models for dimension reduction. Two models are considered. Both model the function with B-splines. The first model assumes an autoregressive model of order 1 for the correlation matrix of the data. The second model considers general between and within-group covariance matrices, with diagonal matrices as a special case, for the coefficients of the splines. The models are tested on microspectrophotometry data for pen inks and for cotton and woollen fibres.

## **RF5: Rapid Fire Talks**

**Wednesday 5 September 10am - 10.50am**

### **Investigation of VAT Expenditure as a quarterly source of Intermediate Consumption**

William Perks, Andrew Sutton

ONS

Objective of the work, This workstream aims to develop an estimate of quarterly intermediate consumption for use in Supply Use balancing within National Accounts. Methods or models used, To best investigate this objective, we have built a prototype VAT expenditure-specific pipeline that outputs industry intermediate consumption. This has been achieved using distributed networks, open source query tools and coding technologies for storing, processing, querying and analysing. We will describe methods used with this pipeline including matching and linking, apportionment, calendarisation, estimation, conceptual adjustments Results or conclusions reached (as appropriate). We will explain comparisons with the Annual Business Survey and Annual Purchases Survey estimates of intermediate consumption at both an aggregate and microdata level including industry IC comparisons. We will also explore comparisons between quarterly VAT expenditure and VAT turnover

## **RF5: Rapid Fire Talks**

**Wednesday 5 September 10am - 10.50am**

### **Estimating heterogeneity variance under sparsity**

Susan Martin, Dankmar Boehning  
*University of Southampton*

Meta-analysis has become the gold standard in medical research analysis in the past few decades. The random-effects model is generally the preferred method to conduct a meta-analysis, as it incorporates between-study heterogeneity - the variability between study estimates as a result of differences in study characteristics. Several methods to estimate the heterogeneity variance parameter in this model have been proposed, including the popular DerSimonian-Laird estimator. Many medical meta-analyses contain few studies, have small sample sizes, or are concerned with rare-event data, where event probabilities are so low that often a small number or zero events are observed in the studies. An example of this is adverse drug reactions in a clinical trial. In such cases, most pre-proposed heterogeneity variance estimators perform poorly, and standard analysis techniques can result in the incorrect estimation of overall treatment effect. We propose some novel methods that we believe are appropriate for the estimation of heterogeneity variance in the case of rare-event data. These are based on generalised linear mixed models (GLMMs), and include the use of the Poisson mixed regression model and the conditional logistic mixed regression model. We are conducting a simulation study to compare our novel GLMM-based techniques with a selection of pre-proposed heterogeneity variance estimators for use in random-effects binary outcome meta-analyses. Our aim is to investigate a variety of realistic scenarios found in sparse-event data, simulating meta-analyses for each scenario, and then determining the performance of the heterogeneity variance estimators in terms of measures such as bias and mean squared error. From the results produced so far in our simulation study, we have found that our novel GLMM-based estimating methods appear to perform well in terms of the estimation of heterogeneity variance with rare-event data, when compared to pre-proposed estimators, especially when study sample sizes in the meta-analysis are highly unbalanced.

## RF5: Rapid Fire Talks

Wednesday 5 September 10am - 10.50am

### **Evaluating service development in critical care: Impact of establishing a medical high dependency unit on intensive care unit workload, case mix, and mortality**

Gordon Prescott,<sup>1</sup> Nabeel Amiruddin<sup>2</sup>, Douglas Coventry<sup>3</sup>, Jan Jansen<sup>4</sup>

<sup>1</sup>*Medical Statistics Team, Institute of Applied Health Sciences, University of Aberdeen,*

<sup>2</sup>*Russell Hall Hospital, Dudley Group Hospitals,* <sup>3</sup>*Department of Intensive Care Medicine, Aberdeen Royal Infirmary,* <sup>4</sup>*Division of Acute Care Surgery, Department of Surgery, University of Alabama*

Background: Critical care services underpin the delivery of many types of secondary care, and there is increasing focus on how to best deliver such services. The aim of this study was to investigate the impact of establishing a medical high dependency unit, in a tertiary referral centre, on the workload, case mix, and mortality of the intensive care unit. Methods: Single-centre, 11-year retrospective study of patients admitted to the general intensive care unit, before and after the opening of the medical high dependency unit, using interrupted time series methodology. Results: Over the duration of the study period, 3209 medical patients were admitted to the intensive care unit. There was a constant rate of medical admissions to the intensive care unit until the opening of the medical high dependency unit, followed by a statistically significant decline thereafter. There was a statistically significant decrease in the average severity of illness of medical patients prior to the opening of the medical high dependency unit, but there was no evidence of a change following the opening of the unit. There was no evidence of a statistically significant change in the estimated mean standardised mortality ratio for either medical or surgical admissions after the intervention. Conclusions: The opening of a medical high dependency unit had a minimal impact on the intensive care unit. There was, in all likelihood, an unmet need - of less seriously ill patients, who were previously looked after on a normal ward, but did not require intensive care unit admission - who are now cared for in the new medical high dependency unit. Interrupted time series analysis, although not without limitations, is a useful means of evaluating changes in service delivery.

## **RF5: Rapid Fire Talks**

**Wednesday 5 September 10am - 10.50am**

### **Perhaps the censoring in my data is not non-informative. Does it matter?**

Alan Kimber, Stefanie Biedermann  
*University of Southampton*

Standard text book survival analyses of censored data are based on the assumption that the censoring mechanism is non-informative. Unfortunately, the usual survival data format - where we have for each observation in the sample the smaller of the survival time and the censoring time, together with an event indicator that tells us whether the time we observe is a survival time or a censoring time - does not allow us to test whether censoring is non-informative or not. So it is of interest to have methods to assess the sensitivity of inferences to the presence of censoring that is not non-informative, thereby allowing us at least to carry out a "what if?" analysis. In this talk we will discuss some methods for achieving this and highlight some difficulties.

## **RF5: Rapid Fire Talks**

**Wednesday 5 September 10am - 10.50am**

### **A case-subcohort study for pancreatic cancer in the China Kadoorie Biobank**

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 designed a case-subcohort study within the China Kadoorie Biobank, a prospective cohort study of over 0.5M 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 on 700 pancreatic cancer cases that accumulated over about 8 years of follow-up and a subcohort of 700 individuals. We used weighted Cox proportional hazards models to assess the associations between metabolic markers and pancreatic cancer risk. Some methodological issues and some preliminary results will be discussed.

## **RF5: Rapid Fire Talks**

**Wednesday 5 September 10am - 10.50am**

### **Women and Children's Health Management Big Data Platform---A Case on National Health and Family Planning Commission of the People's Republic of China**

Huaihai Hui, Des McLernon, Zaidi Sar, Wei Li  
*University of Leeds*

By presenting the Women and Children's Health Management Big Data Platform, I will introduce the construction of the Big Data Analysis Platform of China's National Health and Family Planning Commission and also discuss the latest trends and methods in data statistics. In particular I will examine the following questions: How do we define both Structured and Unstructured Data? How do we get the Population Registration Data from the system of the Ministry of Public Security? How do we get the Medical Institutions Data from the system of the State Administration for Industry and Commerce? How do we get GIS Data from the system of the Ministry of Land and Resources, etc.? At the same time, how does the data provided and analyzed by this Big Data Platform help the medical administrators make fundamental decisions? I will also introduce the various components of the platform: the Computation Engine, the Data Integration and Exchange System, the Storage System, the Analysis and Mining Services, the Platform Operation and the Protection and Maintenance. Finally, I will examine the Data Integration Framework, the Architecture and the Database of this particular Big Data Platform.

## **RF6: Rapid Fire Talks**

**Wednesday 5 September 10am - 10.50am**

### **Grouping anti-social behaviour types based on co-occurrence or impact?**

Puneet Tiwari

*Nottingham Trent University*

Anti-social behaviour (ASB) is a major policing problem in UK. A number of scholars and practitioners have stressed the need to follow an approach different from that dealing with crime when tackling anti-social behaviour. ASB occurs in a variety of form, often hard to distinguish from various low-impact offences. However, dealing with these separately is immense stress on policing resources which are already scarce. In Crime Survey for England and Wales (CSEW), a key national crime-statistics resource, ASB is categorised into 13 different types! Whether ASB should be categorised based on its impact or co-occurrence of these 13 ASBs is a question very relevant to practitioners. This study uses CSEW to provide a justification on a joint approach to categorise ASBs, as both the criteria are important enough to be taken in to account. The study explores various dissimilarity indices and correlation to explore relation between various ASBs. Results show that following any single approach provides very different results. The study then devices an aggregate similarity index with weights equal to the frequency of repeat incidences. This accounts for cases with high impact and low co-occurrences (vulnerable victims), as well as cases with low impact but high incidences.

## **RF6: Rapid Fire Talks**

**Wednesday 5 September 10am - 10.50am**

### **Modelling reassurances of clinicians using Hidden Markov Models**

Valentin Popov

*University of St Andrews*

A key element in the interaction between clinicians and patients with cancer is reassurance giving. Learning about the stochastic nature of reassurances as well as making inferential statements about the influence of covariates such as patient response and time spent on previous reassurances are of particular importance. In this paper we fit Hidden Markov Models (HMMs) to reassurance type from multiple time series of clinicians' reassurances. The data is decoded from audio files of review consultations between patients with breast cancer and their therapeutic radiographer. Assuming a latent state process driving the observations process, HMMs naturally accommodate serial dependence in the data. Extensions to the baseline model such as including covariates as well as allowing for fixed effects for the different clinicians are straightforward to implement. We found that patient's response as well as the duration of the previous reassurance influence the behaviour of clinicians. In particular, clinicians are more likely to switch between the states if the previous reassurance was lengthy. Furthermore, a short reassurance almost certainly leads to maintaining the preference for reassurances that simply instruct the patient not to worry and do not provide informative statements.

## **RF6: Rapid Fire Talks**

**Wednesday 5 September 10am - 10.50am**

### **Correcting unwanted variation in RNA sequencing data derived from a multi-centre study of leukemia**

Anna Quagliari, Terence Paul Speed, Ian Majewski, Edward Chew  
*Walter and Eliza Hall Institute of Medical Research*

Core-Binding Factor Acute Myeloid Leukemia (CBF-AML) is a subtype of AML that represents approximately 15% of all AML cases. Although 90% of patients with CBF-AML attain complete remission after treatment, the disease will return in >40% of cases. We hope to develop a molecular profile that will help predict patient outcome, specifically within the CBF-AML group. We assembled a large cohort of diagnostic CBF-AML RNA Sequencing (RNA-Seq) samples, gathered from our own centre and from public repositories. We are investigating whether gene expression differences, or somatic mutations in key AML genes, have predictive power in CBF-AML. I will present the statistical challenges that we face when combining RNA-Seq data from different labs, different tissues and with variable tumour content. I will show that by exploiting biological and technical replicates, as well as negative control genes (genes not influenced by the factor of interest) we are able to better account for this variability. I will present some initial results on detecting marker genes associated with poor outcome, which we hope will further inform our understanding of this deadly disease.

## **RF6: Rapid Fire Talks**

**Wednesday 5 September 10am - 10.50am**

### **Stratified multivariate permutation tests for mixed variables with applications in industrial statistics**

Luigi Salmaso, Rosa Arboretti, Eleonora Carozzo  
*University of Padova*

In most of the literature the stratified two-sample problem is faced in the field of the so called multiclinic trials. In this presentation we illustrate the problem in a completely different framework beginning from a genuine industrial example and we propose a solution based on permutation tests and nonparametric combination methodology (NPC). Since we handle with multivariate (possibly dependent) variables we compare our proposed approach with a relatively recent solution based on an extension for stratified multivariate data of the well known Mann-Whitney estimator. Furthermore, in order to show the advantage of considering a stratified test instead of a traditional two- sample multivariate test, we also consider as competing method the Hotelling's test. The results of the simulation study show that our test performs better than competitors in several situations and gains power when the number of strata increases.

## RF6: Rapid Fire Talks

Wednesday 5 September 10am - 10.50am

### **MAIC-ing the most of trials?: A Bayesian exploration of Matching Adjusted Indirect Comparison in a Network Meta Analysis**

Joy Leahy<sup>1</sup>, Cathal Walsh<sup>2</sup>

<sup>1</sup>Trinity College Dublin, <sup>2</sup>University College Limerick

Incorporating Individual Patient Data (IPD) into a Network Meta Analysis (NMA) is considered the gold standard of analysis, as it allows a more in-depth analysis of the data, and accounts for differences in covariates between trials. However, the situation can often arise where a researcher has IPD for trials concerning a particular treatment (for example from a sponsor), but none for other trials. In this case, one can re-weight the IPD so that the covariate characteristics in the IPD trials match that of the aggregate data (AD) trials, using a method called Matching Adjusted Indirect Comparison (MAIC). We carry out a simulation study to investigate this method in a Bayesian setting. We simulate 3 IPD trials comparing treatments A and B (AB-IPD trials), and one aggregate data trial comparing treatments B and C (BC-AgD trial). We investigate two options of weighting covariates: 1. all three studies are weighted separately to match the BC-AgD trial (MAIC Separate Trials). 2. patients are weighted across all three IPD studies to match the BC-AgD trial, but the NMA still considers each trial separately (MAIC Pooled Trials). We apply these methods to a network of treatments for multiple myeloma. MAIC can provide more accurate estimates of the relative treatment effects in the BC-AgD trial population. However, MAIC will decrease the accuracy of the relative treatment effects in the overall population. Treatment rankings were unchanged when applying MAIC to the multiple myeloma network. MAIC is beneficial as a sensitivity analysis to confirm results hold across patient populations. If there is a difference in relative treatment effects attributable to population imbalances, then it is useful to be able to quantify this difference. Given the increasing use of MAIC, it is important that researchers think carefully about the population of interest before conducting an MAIC.

## **RF7: Rapid Fire Talks**

**Wednesday 5 September 10am - 10.50am**

### **A systematic review and classification of the methods to `Borrow Strength` in Health Technology Assessment**

Georgios Nikolaidis, Marta Soares, Beth Woods, Stephen Palmer  
*Centre for Health Economics, University of York*

Objective: Sparse relative effectiveness evidence is a relatively common problem in Health Technology Assessment (HTA). For example, evidence on a particular comparator or randomised evidence in a paediatric population may be limited. Where evidence directly pertaining the decision problem is sparse, one option is to expand the evidence-base to include studies that relate to the decision problem only indirectly: for instance, if there is no evidence on a comparator, evidence on other treatments of the same molecular class could be used; similarly, a decision on children may borrow strength from evidence on adults. Usually, such indirect evidence is either included by ignoring any differences (`lumping`) or is not allowed to influence the decision (`splitting`). However, a range of more sophisticated methods exist in the literature which, rather than `lumping` or `splitting`, borrow strength from the indirect evidence while accounting for potential heterogeneity.

Methodology / Classification of methods: We systematically searched the literature for models extending traditional network meta-analysis to allow borrowing strength. We identified 70 papers explaining such methods. The methods were categorised according to three characteristics: 1. The `core` relationship used for borrowing (exchangeability, functional relationships, multivariate relationships and prior-based relationships), 2. The level of the PICOS (Populations, Interventions, Comparators, Outcomes, Study designs) that they included indirectly related evidence on and 3. The parameter they imposed statistical models on (relative treatment effect, bias, between-trials variance etc). The HTA-related methodological issues that the methods were used to address were discussed throughout.

Conclusions: There are often several methods that can be applied to `borrow strength`. These rely on different assumptions and impose varying degrees of strength borrowing. The produced classification of methods according to the `core` relationship, makes the assumptions imposed by each methodology explicit, in an attempt to systematize the choice of methods to synthesize evidence which directly and indirectly relates to a decision.

## **RF7: Rapid Fire Talks**

**Wednesday 5 September 10am - 10.50am**

### **Modeling Three Dimensional T-cell Motility Using Multivariate Normal Distribution**

Elaheh Torkashvand, Joel A. Dubin, Greg Rice  
*University of Waterloo*

Modeling the way immune cells move has received considerable attention recently. Biologists aim to study the immune cells movement patterns when there is no antigen/treatment imposed into collagen. Recently, we have proposed to cluster three-dimensional spatio-temporal trajectories independent of their direction. Homogeneous and non-homogenous hidden Markov models (HHMM and NHMM) were fitted to the angular change of spherical coordinates accordingly. However, due to the label switching problem, we modelled the step length after the local decoding of the fitted Markov chain. In this paper, we propose a transformation of the spherical coordinates. The correlation between different components of transformed spherical coordinates are estimated by fitting HHMM/NHMM where the components follow multivariate normal distributions. We also compare the performance of existing motility models in the prediction of motility metrics of three-dimensional spatio-temporal trajectories by introducing loss functions.

## **RF7: Rapid Fire Talks**

**Wednesday 5 September 10am - 10.50am**

### **Evaluating kidney function while allowing for selection bias**

Matthew Robb

*NHS Blood and Transplant*

The estimated glomerular filtration rate (eGFR) is an important indicator of kidney function following organ transplantation. In a clinical trial to compare two methods of storing a kidney between retrieval from an organ donor and transplantation into the recipient, eGFR at one year is one of the main response variables to be used in treatment comparisons. However, this response variable can only be measured in patients whose kidney graft survives to one year. Therefore, comparing average eGFR between treatments for those whose graft has survived will lead to biased results. One way of accounting for this effect is to use methods developed in econometrics for handling selection bias, such as the Heckman procedure or a full likelihood approach, and this will be illustrated. In order to use the appropriate analyses in the clinical trial, the performance of the resulting estimators must be evaluated under different conditions, for example, when survival and the eGFR both depend on a treatment effect. Here we investigate the use of a likelihood based approach and compare to the Heckman selection method and simply analysing the observed data. Results based on simulated data indicate that comparing observed means often gives reasonable results. However, when there is an important term in both levels of the model and a high level of correlation between the levels, the likelihood method outperforms the other methods.

## RF7: Rapid Fire Talks

Wednesday 5 September 10am - 10.50am

### A detection algorithm for the first jump time in jump-diffusions

Jiao Song,<sup>1</sup> Jiang-Lun Wu<sup>2</sup>

<sup>1</sup>Swansea University Medical School, <sup>2</sup>Department of Mathematics, Swansea University

Background: Beyond modelling uncertainty by Gaussian or normal distributions, non-Gaussian models, which accept asymmetry and/or skewness, could be vital candidates of modelling data featured with heavy-tailed distributions. Such models, which use probability distributions with infinite moments known as Lévy distributions after Paul Lévy, are more realistic than Gaussian models when dealing with heavy tail type data.

Objective: We are interested in the area of developing generalisations of Lévy processes to create new tools to study nonlinear dynamics and kinetics. We present our approach of deriving an algorithm to detect the first jump time in sample trajectories of some jump diffusions.

Method: The two classes of jump diffusions considered are described as solutions of stochastic differential equations (SDEs) driven by  $\alpha$ -stable Lévy motions. One class is the SDEs with linear drift coefficient and additive  $\alpha$ -stable noise and the other one is the linear SDEs, i.e. SDEs with linear drift and diffusion coefficients or the linear SDEs with multiplicative  $\alpha$ -stable noise. We utilise computer simulation algorithm in MATLAB to visualise the trajectories of the jump diffusions for various combinations of parameters arising in the modelling structure of SDEs and adopt a multivariate Lagrange interpolation approach to detect the first jump point.

Conclusion: The algorithm gives a simple yet efficient approach to detect first jump point. Our ongoing work is to extend the algorithm and link to sample data in various scenarios after model identification.

## RF7: Rapid Fire Talks

Wednesday 5 September 10am - 10.50am

### **Effectiveness of a school based smokeless tobacco intervention in Karachi, Pakistan: A cluster randomized trial**

Dr Shafquat Rozi,<sup>1</sup> Nida Zahid<sup>2</sup>, Talat Roome<sup>3</sup>, Maryum Ali<sup>2</sup>, Sobiya Sawani<sup>2</sup>, Anam Razzak<sup>3</sup>, Zahid Ahmad Butt<sup>4</sup>

<sup>1</sup>Department Of Community Health Sciences, Aga Khan University, Karachi, Pakistan, <sup>2</sup>Aga Khan University, <sup>3</sup>Dow University of Health Science., <sup>4</sup>School of Population and Public Health, University of British Columbia, Vancouver, Canada

**Background:** Smokeless tobacco (SLT) use and its consequences such as oral cancers are a significant public health issue globally. We aimed to implement an intervention study in secondary schools to improve the knowledge, attitude and perception regarding betel quid, areca-nut and chewable tobacco use and its association with oral cancers among students in grade 6-10.

**Methods:** A school-based cluster randomized intervention trial was carried out in 18 secondary schools targeting male and female school students from grades 6 to 10 in Karachi. Intervention comprised of health education sessions, posters, pictorial booklets, and a video game on the hazards of use of various tobacco products. Primary outcomes were mean knowledge, attitude and perception scores about hazards of smokeless tobacco use. We performed multivariable analysis using generalizing estimating equation (GEE) to assess the association of various factors with knowledge regarding harmful effects of smokeless tobacco use.

**Results:** We enrolled 738 participants in Intervention group and 589 in the control group. The mean score of knowledge, attitude and perception was significantly improved in intervention group as compared to control group (p value <0.01). Multivariable analysis indicated an increase by 6.34 units (95% CI: 5.54, 7.15) in the mean score of knowledge among intervention as compared to control group. 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, children who were getting information regarding harmful effects of SLT use in school and in textbooks and children who had friends using SLT.

**Conclusions:** Our study indicates that a school-based intervention programme is effective in increasing knowledge regarding the harmful effects of smokeless tobacco use among school going adolescents. Introduction of such educational programmes on regular basis or as part of school curriculum can have an impact on reducing or stopping smokeless tobacco use.

## **Keynote 4 - Significance Tests: Rethinking the Controversy**

**Wednesday 5 September 11.20am – 1.20pm**

Intermingled in today's statistical controversies are some long-standing, but unresolved, disagreements on the nature and principles of statistical methods and the roles for probability in statistical inference and modelling. In reaction to the so-called "replication crisis" in the sciences, some reformers suggest significance tests as a major culprit. To understand the ramifications of the proposed reforms, there is a pressing need for a deeper understanding of the source of the problems in the sciences and a balanced critique of the alternative methods being proposed to supplant significance tests.

### **What do we learn about significance tests from the replication crisis?**

Deborah Mayo

*Department of Philosophy, Virginia Tech*

Today's problems of replication and questionable research practices have implications for the proper interpretation and rationale of significance tests. What are they, and what follows for proposed reforms?

### **Should we redefine statistical significance?**

Richard Morey

*School of Psychology, Cardiff University*

Recently, a team of statisticians and scientists suggested redefining statistical significance to  $p < .005$ , primarily on Bayesian grounds (Benjamin et al, 2017). Their arguments are explored from both Frequentist and Bayesian perspectives; it will be suggested that they cannot be sustained from either.

### **The Use and Abuse of Significance Testing and its Role in Statistical Model Validation**

Aris Spanos

*Department of Economics, Virginia Tech*

Fisher's significance testing is placed in a broader frequentist testing context with a view to shed light on the merits and demerits of its application in several different contexts. It is argued that the current discussions of the use and abuse of the p-value and the Neyman-Pearson (N-P) accept/reject rules as a basis for providing evidence for or against particular hypotheses need to be clarified by drawing three crucial distinctions: (i) the modeling vs. the inference facet of statistical analysis, (ii) testing within and testing outside the boundaries of a statistical model, and (iii) statistical vs. substantive model validation. Focusing on the use of significance testing in statistical model validation, it is argued that it can be modified to provide a sound basis for statistical misspecification testing [testing outside] with a view to establish the adequacy of a statistical model: the validity of its probabilistic assumptions vis-à-vis the particular data. Indeed, a modified form of significance testing can suggest ways to respecify a misspecified statistical model with a view to account for the overlooked systematic statistical information in the data.

## **In Gentle Praise of Significance Tests**

Sir David Cox

*Nuffield College, Oxford*

Four distinct contexts are distinguished and exemplified in which significance tests may be helpful. The connection with general notions about statistical inference is outlined.

## **6.1 Medical: Can clinical prediction be improved with the use of longitudinal data?**

**Wednesday 5 September 2.30pm - 3.50pm**

### **Use of repeated measurements of risk factors in electronic health records to predict future cardiovascular disease risk**

Michael Sweeting,<sup>1</sup> Ellie Paige<sup>2</sup>, Jessica Barrett<sup>3</sup>, David Stevens<sup>3</sup>, Ruth Keogh<sup>4</sup>, Irene Petersen<sup>5</sup>, Angela Wood<sup>3</sup>

<sup>1</sup>*University of Leicester*, <sup>2</sup>*The Australian National University*, <sup>3</sup>*University of Cambridge*,

<sup>4</sup>*London School of Hygiene and Tropical Medicine*, <sup>5</sup>*University College London*

Many prediction models have been developed for the risk assessment and prevention of cardiovascular disease, although most are based on using single measurements of common risk factors. Recently, efforts have focussed on assessing the added value of incorporating repeated measures of common risk predictors. With the benefits of using electronic health records for disease risk screening and personalised healthcare decisions becoming increasingly recognised, such an approach is now more relevant than ever. In this talk I will review statistical methods that have been proposed to incorporate repeat measurements into cardiovascular prognostic models, with the focus on simple methods that are scalable to large primary care databases. Using two-stage landmark approaches, historical repeat measurements can be summarised with cumulative averages or mixed-effects models and incorporated into dynamic prediction models. I will discuss the advantages of using mixed-effects models, which can be extended to estimate multivariate longitudinal profiles and can easily handle sporadically recorded repeat measures and unobserved data. Dynamic risk prediction models are developed using landmark-age models, which provide relevant predictions for age-groups of interest. The methods will be illustrated using data from 41,373 individuals contributing primary care data to The Health Improvement Network.

## **6.1 Medical: Can clinical prediction be improved with the use of longitudinal data?**

**Wednesday 5 September 2.30pm - 3.50pm**

### **Mixed Model tools for multivariate longitudinal classification in R**

David Hughes

*University of Liverpool*

Multivariate mixed models are a useful tool for modelling multiple longitudinal markers. However, these models are complex and estimating model parameters can be complicated. In this talk we briefly survey approaches to fitting multivariate mixed models in R. We then give a demonstration of the use of the R package mixAK to perform the multivariate longitudinal discriminant analysis described in the first talk.

## **6.1 Medical: Can clinical prediction be improved with the use of longitudinal data?**

**Wednesday 5 September 2.30pm - 3.50pm**

### **Improving the accuracy of diagnostic/prognostic models using longitudinal data**

Marta García-Fiñana<sup>1</sup>, David M. Hughes<sup>1</sup>, Arnost Komarek<sup>2</sup>

<sup>1</sup>*University of Liverpool*, <sup>2</sup>*Charles University, Prague*

The use of longitudinally observed data (data collected over time) can be particularly powerful in discriminant analysis to make predictions of the future status of a patient (e.g., whether a patient will develop liver cancer within 5 years). This talk will start with an overview of a recently developed multivariate longitudinal discriminant approach, where longitudinal markers of different types (continuous, counts, binary) are jointly modelled. We will explore the benefits of taking into account the level of uncertainty of group membership probabilities to improve classification. The proposed approach will be illustrated using clinical data.

### **6.3 Applications of Statistics: Ensuring reliability of forensic scientific evidence**

**Wednesday 5 September 2.30pm - 3.50pm**

#### **How the Regulator monitors and influences the quality of scientific evidence**

Jeff Adams

*Home Office*

The presentation shall briefly set out the role of the Regulator and how the Regulator sets, and monitors, quality standards for forensic science. Building on the basic role there shall be consideration of the way in which statistics has featured in the setting of standards and addressing problems.

## 6.3 Applications of Statistics: Ensuring reliability of forensic scientific evidence

Wednesday 5 September 2.30pm - 3.50pm

### Testing for drug driving: determining the measurement error distribution

Simon Spencer,<sup>1</sup> Jeff Adams<sup>2</sup>

<sup>1</sup>*University of Warwick*, <sup>2</sup>*Forensic Science Regulation Unit, Home Office*

Testing blood samples taken from drivers suspected of drug driving is performed using Liquid Chromatography Mass Spectroscopy (LCMS). The procedure is more complex than testing the blood alcohol level due to the fact that samples are tested for multiple substances, there is greater potential for contamination and the analytical technique is more variable. The key challenge is to quantify the measurement uncertainty in a format that can be used during prosecutions. In this work I discuss the analysis of some replicate measurements taken from quality control data in which the true drug concentrations are known in advance. This allows us to examine the distribution of measurement errors, which appears to have some surprising and unexpected properties. Although as a statistician I would like to fit a very complex statistical model to capture the shape of the error distribution, this kind of analysis would have little value to the Forensic Science Regulation Unit because it contains a large number of assumptions and would be difficult to explain in court. Is there a simple alternative that is flexible enough to correctly communicate the measurement uncertainties?

## 6.4 Social Statistics: Multilevel modelling in the social sciences

Wednesday 5 September 2.30pm - 3.50pm

### Local Regularization in Spatially-Correlated Multilevel Models

Levi Wolf

*University of Bristol*

Multilevel (or variance components) models have been applied in many areas of regional science, epidemiology and polimetrics. They are most often used to model treatment nonstationarity in policy regimes, a form of spatial process heterogeneity. Multilevel models with spatially-correlated components are increasingly used to model the presence of both spatial heterogeneity and spatial dependence. Previous treatments typically focus on studying a single model specification in detail, deriving its estimators, and demonstrating its properties in a case study. Instead, a general approach is possible. In this paper, a generic Gibbs sampler for spatially-correlated multilevel models is developed and its empirical properties examined in model of cancer screening. Critically, there is tension between spatial dependence and spatial heterogeneity terms in the model. This results in apparent "growth" in some multilevel models where the dependence terms overpower the effects of estimate regularization. This results in locally-smoothed estimates, which, depending on the strength of spatial dependence, may be both larger and more precise than their corresponding single-level standard linear models. Thus, shrinkage always applies, but not always from the standard single-level linear model. Due to the complexity of these tradeoffs between model fit, dependence, and heterogeneity terms, this formal attention is critical to building systematic and transferable understanding.

## **6.4 Social Statistics: Multilevel modelling in the social sciences**

**Wednesday 5 September 2.30pm - 3.50pm**

### **Avoiding Bias When Estimating the Consistency and Stability of Value-Added School Effects Using Multilevel Models**

George Leckie

*University of Bristol*

The traditional approach to estimating the consistency of school effects across subject areas and the stability of school effects across time is to fit separate value-added multilevel models to each subject or cohort and to correlate the resulting empirical Bayes predictions. We show that this gives biased correlations and these biases cannot be avoided by simply correlating “unshrunk” or “reinflated” versions of these predicted random effects. In contrast, we show that fitting a joint value-added multilevel multivariate response model simultaneously to all subjects or cohorts directly gives unbiased estimates of the correlations of interest. There is no need to correlate the resulting empirical Bayes predictions and indeed we show that this should again be avoided as the resulting correlations are also biased. We illustrate our arguments with separate applications to measuring the consistency and stability of school effects in primary and secondary school settings. However, our arguments apply more generally to other areas of application where researchers routinely interpret correlations between predicted random effects rather than estimating and interpreting these correlations directly.

## **6.4 Social Statistics: Multilevel modelling in the social sciences**

**Wednesday 5 September 2.30pm - 3.50pm**

### **Automating Multilevel analysis and statistical teaching preparation**

William Browne

*University of Bristol*

Computer software has for many years been an integral part of a statisticians daily work and doing statistical analysis can be thought of as a partnership between the statistician and the software. In this talk we look at how much of this combined work the computer might in theory do. In our recent ESRC funded research we have created within our StatJR software functionality that creates Statistical Analysis assistants which are software programs that with limited user input will perform a complete statistical analysis. We will talk about how this allows (semi) automated multilevel analysis. We have also in British Academy funded work added further automation functionality to StatJR with this time the aim of automating the production of teaching materials. Here we have worked with the SPSS package and StatJR will construct bespoke training materials using a researchers own dataset. We will describe these features and consider how these features might benefit the statistics community in the future.

## **6.5 Methods & Theory: Recent developments in covariance modelling**

**Wednesday 5 September 2.30pm - 3.50pm**

### **Joint modelling of survival and longitudinal data**

Hongsheng Dai

*University of Essex*

For joint modelling of survival and longitudinal data, the two submodels are often linked via random effects. We will present a nonparametric approach for dealing with such unknown random effects and an asymptotically consistent working likelihood is used in our parameter estimation. This new method is presented with an application to a primary biliary cirrhosis study, where informative observation times are involved.

## 6.5 Methods & Theory: Recent developments in covariance modelling

Wednesday 5 September 2.30pm - 3.50pm

### A calibration method for non-positive definite covariance matrix in multivariate data analysis

Chao Huang,<sup>1</sup> Jianxin Pan<sup>2</sup>, Daniel Farewell<sup>3</sup>

<sup>1</sup>University of Hull, <sup>2</sup>University of Manchester, <sup>3</sup>Cardiff University

This presentation relates to the invited session on covariance modelling. Background: The importance of correctly modelling the correlation and covariance structure in multivariate data analysis has been more and more recognized. Covariance matrices that fail to be positive definite arise often in covariance estimation. One typical example is the sample variance, which is often singular when the sample size is close to, or less than, the dimension of the random samples. Methods: In this work, we propose a unified statistical and numerical matrix calibration, finding the optimal positive definite surrogate in the sense of Frobenius norm. Our proposed approach is implemented through a straightforward screening algorithm. Results: Simulation results showed that the calibrated matrix is typically closer to the true covariance, while making only limited changes to the original covariance structure. We also revisited two substantive analyses to demonstrate the properties of the proposed calibration. Conclusion: This approach is not constrained by model assumptions. Neither is it limited by data structures. Since it is a calibration approach, it can be incorporated in existing covariance estimation process, and offers a routine check and calibration of covariance matrix estimators. One R package was also built for spreading the usage of this approach.

## **6.5 Methods & Theory: Recent developments in covariance modelling**

**Wednesday 5 September 2.30pm - 3.50pm**

### **Joint Mean-covariance Modelling and its R Package: JMCM**

Yi Pan

*University of Birmingham*

Longitudinal studies are commonly arising in various fields such as psychology, social science, economics and medical research, etc. It is of great importance to understand the dynamics in the mean function, covariance and/or correlation matrices of repeated measurements. However, the high-dimensionality (HD) and positive-definiteness (PD) constraints are two major stumbling blocks in modelling of covariance and correlation matrices. It is evident that Cholesky-type decomposition based methods are effective in dealing with HD and PD problems, but those methods were not implemented in statistical software yet, causing a difficulty for practitioners to use. In this talk, I will introduce three Cholesky decomposition based methods for joint modelling of mean and covariance structures, namely Modified Cholesky decomposition (MCD), Alternative Cholesky decomposition (ACD) and Hyperspherical parameterization of Cholesky factor (HPC). I will then introduce our newly developed R package `jmcm` which includes the MCD, ACD and HPC methods. Demonstration will be made by running the package `jmcm` and comparison of those methods will be made through analysing two real data sets.

## **6.8 Industry & Finance: Papers from the Journal of the Royal Statistical Society**

**Wednesday 5 September 2.30pm - 3.50pm**

### **The Repayment of Unsecured Debt by European Households**

Charles Grant

*Brunel University*

The existing literature that estimates the incidence of arrears relies on either household survey data or administrative data derived from the lender's records of their borrowers. But estimates based on these different sources will give different estimates of arrears. Moreover, the estimates are not useful for policy analysis or for the bank's lending decision, since they ignore the fact some households do not borrow. This paper discusses the selection issues involved in using either data source, and is the first paper to bound the estimate of the household's underlying propensity to repay. To demonstrate the methodology, it uses data from the EU-SILC survey for 2008 to estimate the factors that affect repayment among Eurozone households.

## **6.8 Industry & Finance: Papers from the Journal of the Royal Statistical Society**

**Wednesday 5 September 2.30pm - 3.50pm**

### **Do debit cards decrease cash demand? Causal inference and sensitivity analysis using Principal Stratification**

Andrea Mercatanti

*LISER, Luxembourg Institute of Socio-Economic Research*

It has been argued that the use of debit cards may modify the cash holding behaviour, as debit card holders may either withdraw cash from ATMs (Automated Teller Machine) or purchase items using POS (Point of Sale) devices at retailers. In this paper, within the Rubin Causal Model, we investigate the causal effects of the use of debit cards on the cash inventories held by households using data from the Italy Survey of Household Income and Wealth. We adopt the principal stratification approach to incorporate the share of debit card holders who do not use this payment instrument. We use a regression model with the propensity score as the single predictor to adjust for the imbalance in observed covariates. We further develop a sensitivity analysis approach to assess the sensitivity of the proposed model to violation to the key unconfoundedness assumption. Our empirical results suggest statistically significant negative effects of debit cards on the household cash level in Italy.

## **6.8 Industry & Finance: Papers from the Journal of the Royal Statistical Society**

**Wednesday 5 September 2.30pm - 3.50pm**

### **Parameter stability and semiparametric inference in time varying auto-regressive conditional heteroscedasticity models**

Lionel Truquet

*ENSAI*

We develop a complete methodology for detecting time varying or non-time-varying parameters in auto-regressive conditional heteroscedasticity (ARCH) processes. For this, we estimate and test various semiparametric versions of time varying ARCH models which include two well-known non-stationary ARCH-type models introduced in the econometrics literature. Using kernel estimation, we show that non-time-varying parameters can be estimated at the usual parametric rate of convergence and, for Gaussian noise, we construct estimates that are asymptotically efficient in a semiparametric sense. Then we introduce two statistical tests which can be used for detecting non-time-varying parameters or for testing the second-order dynamics. An information criterion for selecting the number of lags is also provided. We illustrate our methodology with several real data sets and obtain some conclusions quite different from the standard fitting obtained with stationary ARCH models.

## 7.1 Contributed - Medical: Joint Modelling and Dynamic Prediction

Wednesday 5 September 4.20pm - 5.20pm

### Dynamic personalised prediction of survival using routinely collected data: An empirical comparison of landmarking and joint modelling

Ruth Keogh,<sup>1</sup> Rhonda Szcznesiak<sup>2</sup>

<sup>1</sup>*London School of Hygiene and Tropical Medicine*, <sup>2</sup>*Cincinnati Children's Hospital Medical Center and University of Cincinnati*

In 'dynamic' prediction of survival we make updated predictions of individuals' survival as new longitudinal measures of health status become available. Dynamic predictions inform patients and clinicians about prognosis based on their data up to the time of the prediction, which can inform treatment decisions and provide personalised information for patients. Two main approaches have been described for obtaining dynamic prediction models: joint modelling of longitudinal and survival data, and landmarking. Recent comparisons of joint modelling and landmarking using simulation studies have tended to find joint modelling to perform slightly better. However, they have focused on simulation scenarios favouring the joint model and have not incorporated practical complications faced in real applications. I will compare the advantages and disadvantages of landmarking and joint modelling via an application in Cystic fibrosis (CF), which is an inherited, chronic, and progressive condition affecting over 70,000 people worldwide. We used data from the US CF Foundation patient registry, which collects longitudinal data on over 29,000 patients, to develop dynamic survival prediction models for 2-, 5- and 10-year survival using both landmarking and joint modelling. Predictors include lung function, weight, infections status, and demographic data. Challenges faced in these data include that some patients receive a transplant; that data are collected from each visit to the clinic, so that frequency of measurements is likely to be informative; that there are several time-dependent predictors of different types; and that we wished to investigate whether variability in lung function over time predicts survival in addition to the absolute level. I will argue that landmarking can perform as well as or better than joint modelling in some circumstances and that landmarking enables us to handle practical challenges in a straightforward way.

## 7.1 Contributed - Medical: Joint Modelling and Dynamic Prediction

Wednesday 5 September 4.20pm - 5.20pm

### Dynamic prediction of conception in couples with unexplained infertility

David McLernon, Amanda Lee, Siladitya Bhattacharya  
*University of Aberdeen*

**Objective**To develop a dynamic prediction model to estimate the chance of conception (leading to the livebirth) from the point of infertility diagnosis and to update this chance at each subsequent month.  
**Methods**Clinical data from 1330 couples with unexplained infertility who attended Aberdeen Fertility Clinic from 1998-2011 were record-linked to national maternity data to obtain pregnancy outcomes. A dynamic Cox regression model was developed using a landmarking approach to predict conception within six months from the point of infertility diagnosis. Predictions were updated monthly up to 22 months. Predictors included female age (fitted as a two-piece linear function), duration of infertility and pregnancy history at baseline, fertility treatment status in each month (clomifene, intra-uterine insemination (IUI), in-vitro fertilisation (IVF), expectant management i.e. no treatment) and month of prediction as a quadratic term. The model was externally validated using a prospective cohort of 5184 patients from The Netherlands who were followed to natural conception. To assess model discrimination, the concordance statistic was calculated at each month of prediction. The calibration slope was used to test for overfitting.  
**Results**The chance of conception declined from the age of 32 and with increasing duration of infertility. On average across all time points, couples who had IUI or IVF had an increased chance of conception compared to expectant management (IUI, Hazard Ratio=2.80 (1.99 to 3.94); IVF, Hazard Ratio=5.02 (4.00 to 6.31)). The concordance statistic of the model applied to the Dutch cohort ranged from 0.56 to 0.67. The calibration slope showed no evidence of overfitting (slope=0.97 (95% CI 0.93 to 1.00)). Dynamic predictions of conception will be presented for couples with different characteristics.  
**Conclusions**This model estimates the diminishing chances of conception with and without treatment over a fixed time horizon. It will inform couples and their clinicians of their estimated chance of success which may help manage expectations throughout their fertility journey.

## 7.1 Contributed - Medical: Joint Modelling and Dynamic Prediction

Wednesday 5 September 4.20pm - 5.20pm

### **Beyond the mean: a joint modelling approach to relate within-individual variation in repeated measures, longitudinal data, to a future outcome**

Richard Parker,<sup>1</sup> Harvey Goldstein<sup>1</sup>, Jon Heron<sup>1</sup>, Laura Howe<sup>1</sup>, George Leckie<sup>1</sup>, Graciela Muniz-Terrera<sup>2</sup>, Kate Tilling<sup>1</sup>

<sup>1</sup>*University of Bristol*, <sup>2</sup>*University of Edinburgh*

Within-individual variability, as well as mean level, of biomarkers may predict later outcomes. For example, within-individual variability in blood pressure (BP) is an independent cardiovascular risk factor above and beyond mean BP (Rothwell, 2010). Naive methods typically used to relate within-individual variability to a distal outcome first calculate a summary measure of variation for each individual – such as the standard deviation or coefficient of variation – and then regress the distal outcome of interest on this summary measure. Such methods have limitations: they do not take account of changes in within-person variability over time; they do not pool information across the sample to improve predictions; and no account is taken of the precision with which the within-individual variation was estimated. We propose a joint modelling approach in which the repeatedly-measured outcome is simultaneously modelled alongside the distal outcome. In the resulting multilevel model, we allow the within-individual variance to be a function of relevant covariates (thus allowing for heteroscedasticity) as well as of a random effect which allows each individual to have more, or less, variation than that implied by the other covariates in the model (Goldstein et al., 2017). This random effect can be readily related to the distal outcome within the same model, for instance by including it in a covariance matrix alongside the residual variance from the distal outcome, or as a predictor for that outcome. We investigate this approach using simulated datasets, and also with data from the Avon Longitudinal Study of Parents and Children (ALSPAC) cohort study: relating variability in blood pressure measurements to later indicators of cardiovascular health. We describe and illustrate estimating these models in a Bayesian framework, sampling the posterior via MCMC methods, and conclude by discussing options for improving the efficiency of model estimation.

## 7.10 RSS Prize Winners - Best presentations of RSC 2018

Wednesday 5 September 4.20pm - 5.20pm

### **Using Health Economics in the Design, Monitoring and Analysis of Adaptive Clinical Trials**

Laura Flight,<sup>1</sup> Steven Julious<sup>1</sup>, Alan Brennan<sup>1</sup>, Susan Todd<sup>2</sup>

<sup>1</sup>*University of Sheffield*, <sup>2</sup>*University of Reading*

Adaptive designs use data collected as a trial progresses to inform modifications to the trial, without compromising validity or integrity. These designs offer an alternative to the traditional 'fixed' sample size design trial where the data are not formally examined until the trial has ended. Adaptive designs have the potential to directly benefit patients and healthcare providers ethically and financially. Adaptive trials are commonly designed such that interim decisions and modifications are based on clinical effectiveness. Despite its importance cost-effectiveness is often only secondary to clinical outcomes. It is currently unclear what impact adaptive designs have on health economic analyses. Additionally, opportunities are potentially being missed to incorporate health economics into the adaptive trial at the design, monitoring and analysis stages. This research aims to explore whether adaptive designs and health economics can be used together to increase the efficiency of health technology assessments without compromising accuracy.

## 7.10 RSS Prize Winners - Best presentations of RSC 2018

Wednesday 5 September 4.20pm - 5.20pm

### **Bayesian Nonparametric Methods for Stochastic Epidemic Models**

Rowland Seymour, Theodore Kypraios, Philip O'Neill  
*University of Nottingham*

Simulating from and making inference for stochastic epidemic models are key strategies for understanding and controlling the spread of infectious diseases. Despite the enormous attention given to methods for parameter estimation, there has been relatively little activity in the area of nonparametric inference. That is, drawing inference for the infection rate without making specific modelling assumptions about its functional form. In this talk we fit heterogeneously mixing models in which the infection rate between two individuals is a function of their characteristics, for example location or type. We develop a novel method for inferring the function nonparametrically, removing the need to make questionable parametric assumptions. We adopt a Bayesian approach by assigning a Gaussian Process (GP) prior to the infection rate function and then develop an efficient data augmentation Markov Chain Monte Carlo methodology to estimate the infection rate function, the GP hyperparameters and the unobserved infection times. We illustrate our methodology using simulated data and by analysing a data set on Avian Influenza from the Netherlands.

## 7.10 RSS Prize Winners - Best presentations of RSC 2018

Wednesday 5 September 4.20pm - 5.20pm

### Parallelisation of a Common Changepoint Detection Method

Samuel Tickle

*Lancaster University*

Dynamic Programming techniques have been a popular means of detecting changepoints in data streams for some time, with methods such as Optimal Partitioning being a popular alternative to Binary Segmentation. The PELT algorithm, based on Optimal Partitioning, reduces the computational cost for a univariate stream from quadratic to linear for most cases, while retaining exactness with respect to the optimality of a penalised cost function. However, in certain cases, PELT remains quadratic, motivating the need for parallel computing to streamline PELT's execution; here, we introduce Chunk and Deal, two means of doing just this, and establish new results on the asymptotic and finite sample properties of the penalised cost function approach to discuss the consistency and computational cost of these approaches under certain conditions. In addition, we shall briefly discuss new results into the extension of the penalised cost function approach to multivariate data streams.

## **7.2 Contributed - Official Statistics & Public Policy: Innovative approaches**

**Wednesday 5 September 4.20pm - 5.20pm**

### **Investigating the feasibility of using alternative data sources in official consumer price statistics**

Jack Philips

ONS

New data sources in the UK consumer price statistics Over the last 5 years, the ONS has started work on exploring new data sources for use in consumer prices statistics. An initial research project ended in summer 2017, with the work being summarised in Research indices using web scraped price data: August 2017 update. Following on from this project, the ONS has recently produced an alternative data sources data collection strategy. This targets the implementation of new data sources in the UK's consumer price statistics by 2020. These data sources include transaction data from retailers and online price data from web scrapers and APIs. This presentation will cover the ONS's strategy and provide an update on the ongoing work that ONS is doing to implement this project. For example, we have been investigating the use of web scraped data to replace some of our online collection of prices that are currently run manually. This includes the collection of data required for our hedonic models. The presentation will cover some of the processes that can be used to clean and validate the data before it can be used in production. Other work streams, such as our methodology research into new indices, will also be covered.

## 7.2 Contributed - Official Statistics & Public Policy: Innovative approaches

Wednesday 5 September 4.20pm - 5.20pm

### **New methodologies for the measurement of inflation using web scraped data in a UK context**

Matthew Mayhew

*Office For National Statistics*

The Office for National Statistics is researching alternative data sources to calculate the Consumer Prices Index including Owner Occupiers' Housing Costs (CPIH). One alternative data source is scraping prices from websites. This has the potential to improve the CPIH in terms of frequency and product coverage, but it also has its complications. Web scraped data has more product churn (products go in and out of stock more frequently), which means that traditional methods of calculation have problems when there are a lot of missing data, as individual products are tracked across time. New more innovative methods need to be developed. There have been many different methods proposed which better measure inflation in this data: Chained Bilateral Jevons – more frequent chaining of the indices to account for the change in samples; Clustering Large datasets into Price indices (CLIP) – tracking groups/clusters of products over time instead of individual products; GEKS-J – Comparing more than two periods in the calculation, allows for new and disappearing items to be included more quickly; Fixed Effects index with a Window Splice (FEWS) – a model based estimator using hedonic models to adjust for the quality change when the set of products changes. Each of these estimators of inflation have their advantages and disadvantage. Their properties have been assessed according to the Axiomatic, Economic and Statistical approaches to Index Numbers. Data for different consumption sectors of the UK market have been used to complete this assessment and the most appropriate estimator for each consumption sector has been proposed.

## **7.2 Contributed - Official Statistics & Public Policy: Innovative approaches**

**Wednesday 5 September 4.20pm - 5.20pm**

### **The feasibility of measuring the sharing economy of the UK**

Pauline Beck, Natalie Jones

*Office for National Statistics*

The sharing economy is widely understood as being the sharing of spare assets or skills through digital means, however there is currently no internationally agreed definition. In this latest paper on the feasibility of measuring the sharing economy, ONS proposes a working definition of the UK sharing economy and a framework for identifying its businesses and users. This conceptual framework supports the collection and dissemination of statistics on sharing economy activities. As a result, new descriptive statistics on business data also shed a light on characteristics of sharing-economy businesses, such as employment costs, advertising and marketing, and turnover. This paper provides new estimates on individuals who use an intermediary app or website to book accommodation or transport from another individual. ONS has also undertaken a data science project to see if it is possible to systematically identify sharing economy companies using variables such as birth date, turnover, Standard Industrial Classification (SIC) and employment. Together, these data provide an invaluable insight into UK sharing economy activity.

### 7.3 Contributed - Applications of Statistics: Medical Applications

Wednesday 5 September 4.20pm - 5.20pm

#### Modification of Cramer-von Mises test with applications in genomics

Alison Telford,<sup>1</sup> Charles Taylor<sup>1</sup>, Henry Wood<sup>2</sup>, Arief Gusnanto<sup>1</sup>

<sup>1</sup>University of Leeds, <sup>2</sup>Leeds Institute for Cancer and Pathology

We propose a novel nonparametric test to identify differences between two different clinical groups. The test is a modification of the Cramer-von Mises test and compares the cumulative distribution estimates. This is motivated by our study on copy number alterations (CNA). CNA is a structural variation in human genome where some regions have more or less copy number than the normal two copies. CNA patterns in some genomic regions across patients have been shown to be associated with disease phenotypes. Our interest is in testing which genomic regions exhibit different distributions between two clinical groups to discover new genomic markers for phenotypic identification. Standard statistical tests, including Cramer-von Mises test, are not adequate to deal with the characteristics of the data where the differences between the two groups lie in the following aspects of the distribution: mean, variance, skewness, and multi-modality. We modify the Cramer-von Mises test by considering a weight function that is anti-proportional to the density function of the pooled data. The results indicate that our proposed method is comparable to other specific tests and preferable to the Cramer-von Mises test when identifying differences in distributions which are multi-modal.

### 7.3 Contributed - Applications of Statistics: Medical Applications

Wednesday 5 September 4.20pm - 5.20pm

#### **A Bayesian statistical analysis to establish best practices for organs-on-chips experiments**

Beate Ehrhardt<sup>1</sup>, Sam Peel<sup>1</sup>, Adam Corrigan<sup>1</sup>, Kyung-Jin Jang<sup>2</sup>, Pedro Pinto<sup>1</sup>, Matt Boeckeler<sup>1</sup>, Lorna Ewart<sup>1</sup>  
<sup>1</sup>*AstraZeneca*, <sup>2</sup>*Emulate*

Microphysiological systems (MPS), or organs-on-chips, emulate physiology at a small scale by engineering appropriate cellular microenvironments. MPS are high content models and our novel automated imaging workflow enables us to robustly capture multi-cellular phenotypes at a high throughput. However, due to the novelty of the technology, there are no best practice standards to analyse the data, determine the sources of variability, or to perform sample size calculations. We have established an analysis pipeline for organs-on-chips to reduce bias and variability. This provides for the first time, a framework of statistical best practice for organs-on-chips experiments. We first use principles of optimal experimental design to randomise the chips in an optimal order. We then analyse the relationship between the fluorescent intensities and treatment and time effects while controlling for the chip-to-chip variability, and potential row or holder effects. As a result, we are able to run a power analysis enabling us to identify the minimum sample size necessary to detect a given effect size. Specifically, we fit a Bayesian multilevel linear regression model with uninformative priors where we treat the fluorescence signal of the fields of view of a chip as repeated measurements. We estimate treatment and time effects as well as the variability induced by chip, row and holder. We demonstrate the functionality of this analysis pipeline using human liver chips by investigating the liver toxicity of a novel AZ compound and contrasting it to the results of a positive control. The analysis pipeline presented here improves the reproducibility of MPS, an important step towards building confidence in the technology.

### 7.3 Contributed - Applications of Statistics: Medical Applications

Wednesday 5 September 4.20pm - 5.20pm

#### Evaluating Community-Based Translational Interventions Using Historical Controls: Propensity Score vs. Disease Risk Score Approach

Luohua Jiang<sup>1</sup>, Janette Beals<sup>2</sup>, Ann Bullock<sup>3</sup>, Spero Manson<sup>2</sup>, Shuan Chen<sup>4</sup>

<sup>1</sup>University of California Irvine, <sup>2</sup>University of Colorado Anschutz Medical Campus, USA,

<sup>3</sup>Indian Health Service, USA, <sup>4</sup>University of California Davis

**Objective:** Due to ethical and other considerations, many community-based translations of evidence-based interventions are designed as one-arm studies. The evaluation of the effectiveness of such programs is challenging. In this study, we evaluate the effectiveness of a translational intervention using historical control data from a publicly available data repository.

**Methods:** Inference based on historical controls could be subject to strong selection bias and imbalance in observed confounders between treatment conditions. We compared the use of propensity scores (PS) and disease risk scores (DRS) to adjust for potential confounder imbalance between groups. These methods were applied to the data from the Special Diabetes Program for Indians Diabetes Prevention (SDPI-DP) demonstration project, a translational lifestyle intervention among American Indian and Alaska Native communities. Publicly available Diabetes Prevention Program (DPP) data was used as a historical control in the evaluation. A newly proposed “dry-run” analysis was employed to evaluate confounding control of the DRS matching approach.

**Results:** The unadjusted hazard ratio (HR) for diabetes risk was 0.35 (95% CI: 0.26-0.43) for SDPI-DP lifestyle intervention vs. control. However, when relevant diabetes risk factors were considered, the adjusted HR estimates were closer to 1 than the unadjusted HR, ranging from 0.56 (95% CI: 0.44-0.71) to 0.69 (95% CI: 0.56-0.96). The differences in estimated HRs using the PS and DRS approaches were relatively small but DRS matching resulted in more participants being matched and smaller standard errors of effect estimates. The “dry-run” analysis results showed adequate confounding control through DRS matching.

**Conclusions:** Carefully employed, publicly available randomized clinical trial data can be used as a historical control to evaluate the intervention effectiveness of one-arm community translational initiatives. It is critical to use a proper statistical method to balance the distributions of potential confounders between comparison groups in this kind of evaluations.

## **7.4 Contributed - Social Statistics**

**Wednesday 5 September 4.20pm - 5.20pm**

### **An Alternative Definition and Bayesian Estimation of Child Labour in India**

Jihye Kim,

*University of Manchester*

Child labour in India involves the largest number of children in the world. This number is estimated to be 11.7 million children aged between 5 and 17 in 2011 according to the latest Indian Census. However, the number seems underestimated due to the narrow definition of child labour used by the Census because it excludes children participating in household activities. Considering the widespread use of child labour in the domestic sector in India, this study suggests extending the definition of child labour taking into account the amount of time spent working at home. How to measure the magnitude of child labour varies according to divergent opinions across international agencies such as ILO and UNICEF. In this study, I use the ILO's methodology to define hazardousness of work and the UNICEF's time threshold for domestic work. The specific aims of the study are first to estimate the prevalence of child labour in the age group 5 to 17 and secondly to combine information from multiple data sources in a Bayesian model to improve the estimation of child labour. This study uses the most recent National Sample Survey on Employment and Unemployment (2011/12) and the India Human Development Survey (2011/12), comparing and combining them with the reported figure of child labour from the Indian Census 2011. This study shows that the number of child labourers (ages 5-17) is estimated higher than the Indian Census, at around 16 million in 2011. The model provides a way to reduce the measurement error due to the use of a single dataset. This method also smooths the variation between ages and provides more reliable estimates of child labour.

## 7.4 Contributed - Social Statistics

Wednesday 5 September 4.20pm - 5.20pm

### Fifty years of multidimensional poverty research in Britain

David Gordon,<sup>1</sup> Hector Najera<sup>1</sup>, Joanna Mack<sup>1</sup>, Marco Pomati<sup>2</sup>, Shailen Nandy<sup>2</sup>, Anne-Catherine Guio<sup>3</sup>, Stewart Lansley<sup>1</sup>

<sup>1</sup>University of Bristol, <sup>2</sup>University of Cardiff, <sup>3</sup>Luxembourg Institute of Socio-Economic Research

Every decade or so since the late 1960s, UK social scientists have attempted to carry out an independent poverty survey to test out new ideas and incorporate current state of the art methods into UK poverty research. Thus, the 1968-69 Poverty in the UK survey (Townsend and colleagues), the 1983 Poor Britain and 1990 Breadline Britain surveys (Mack, Lansley and colleagues) and the 1999 Poverty and Social Exclusion Survey (Bradshaw and colleagues) and its 2002 counterpart in Northern Ireland (Hillyard and colleagues) and the 2012 Poverty and Social Exclusion in the UK survey (Gordon and colleagues) introduced new methods, ideas and techniques about poverty measurement and helped to keep UK academic research at the forefront of poverty measurement methodology. These academic surveys had a considerable policy and methodological impact in the UK, Europe and other countries, even though this was not their primary purpose. For example, the European Union adopted a relative definition of poverty derived from the 1968-69 Poverty in the UK survey. In the 2010 Child Poverty Act the UK Government adopted a combined low income and material deprivation child poverty indicator developed from the 1999 PSE survey. A range of other countries (eg Mexico, New Zealand, Tonga) have adopted similar measures and these methods have been used by numerous academic studies across the world. Finally, the European Union recently adopted a new official measure of Material and Social Deprivation (MSD) and its first Child Deprivation measure developed from the work of the 2012 PSE survey. This presentation will discuss the changes in multidimensional poverty and measurement methods over the past 50 years in the UK. It will also describe how the analytical framework developed by the Poverty and Social Exclusion research projects has been adopted by the European Union to produce suitable, valid, reliable and additive deprivation indicators which are comparable across the 28 EU member countries.

#### 7.4 Contributed - Social Statistics

Wednesday 5 September 4.20pm - 5.20pm

##### **Assessing the role of the field of study in determining university attractiveness: a comparison study based on generalized mixed-effect models**

Silvia Columbu, Isabella Sulis, Mariano Porcu  
*University of Cagliari*

Understanding the phenomenon of intra-and international students mobility has become of increasing relevance in the organization of tertiary education systems. Using information provided by the Italian National Student Archive (NSA) on a cohort of students enrolled for the first time at the university in a.a. 2014/15, we investigate upon the factors that influence students in deciding to attend bachelor degree studies outside the region of residence. The aim of the analysis is twofold: (i) to suggest value-added measures of university and degree program attractiveness and (ii) to assess and split the role played by the field of study of the degree program in determining the power to attract students from other regions. In this perspective we model the probability that freshmen in a given university are mover students (instead that stayer) as a function of their socio-demographic characteristics, territorial area information and other sources of heterogeneity that concern both the field of specialization and the university. This is done by the adoption of a three-level logistic regression model that enables to separate the between-university variability from the between-degree program within-university variability. The NSA data show a cross-classified structure that is analyzed using three different multilevel modelling approaches: (a) a pure cross-classified model where students are clustered in class of degree programs (Level-2) that belong to different universities (Level-3); (b) a hierarchical model that considers that degree programs belonging to the same university - namely degree program-university combinations - (Level-2) are clustered in Universities (Level-3) ; (c) a hierarchical model in which the Level-2 units are defined as in model (b), but that considers that degree programs are clustered in classes of degree (Level-3) (e.g. medicine, statistics, mathematics) which have the same vocation. The three models give different insights for studying the determinants of students' choices and enable to make comparisons between universities and degree programs attractiveness.

## 7.5 Contributed - Methods & Theory: High-dimensional data

Wednesday 5 September 4.20pm - 5.20pm

### On Ridge Estimation for Probability Densities of Dependent Multivariate Data

Jan Beran, Klaus Telkmann

*Department of Mathematics and Statistics, University of Konstanz, Germany*

Understanding geometric and topological features of high-dimensional observations is a key topic for Big Data in medical imaging, phylogenetics, cosmology and others. One of the approaches that can be used for identifying topological features such as holes, tunnels, connectedness etc. is ridge estimation for density functions. In image analysis for instance, statistical inference for the ridge is often the main objective. Formally, in an  $m$ -dimensional space, a  $k$ -dimensional ridge ( $k \leq m$ ) is the set of all points that are local maxima of the density in at least  $(m-k)$  directions. So, for  $k=0$  we have the usual local maxima, for  $k=1$  one gets one-dimensional curves or sets of curves, and so on. In a general setting, our observations will be some high-dimensional point cloud, where the points are generated randomly over time. When there is a ridge, after some time, there will be a concentration of the points around the ridge. Our particular focus is on the situation when there is strong dependence in time. For estimating a ridge, we start with nonparametric kernel density estimators and construct simultaneous confidence sets for density ridges. The formulas are derived by exploiting the so called multivariate reduction principle. For illustrating our methods, we use simulations (programmed in R) in the bivariate case, with circular and spiral ridges as particular examples. This is joint work with Klaus Telkmann.

## 7.5 Contributed - Methods & Theory: High-dimensional data

Wednesday 5 September 4.20pm - 5.20pm

### Gene Hunting with Hidden Markov Model Knockoffs

Matteo Sesia, Chiara Sabatti, Emmanuel Candes  
*Stanford University*

Modern scientific studies often require the identification of a subset of relevant explanatory variables, in the attempt to understand an interesting phenomenon. Several statistical methods have been developed to automate this task, but only recently was the framework of knockoffs proposed as a general solution that can perform variable selection under rigorous type-I error control, without relying on strong modeling assumptions. In this paper, we extend the methodology of knockoffs to a rich family of problems where the distribution of the covariates can be described by a hidden Markov model. We develop an exact and efficient algorithm to sample knockoff variables in this setting and then argue that, combined with the existing selective framework, this provides a natural and powerful tool for performing principled inference in genome-wide association studies with guaranteed false discovery rate control. Finally, we apply our methodology to datasets on Crohn's disease and some continuous phenotypes, e.g. levels of cholesterol. The results of our analysis on datasets from 5 genome-wide association studies show that our method makes more discoveries than its traditional counterparts based on marginal testing, while also offering more easily interpretable results and statistical guarantees based on mild and scientifically sound assumptions.

## **7.5 Contributed - Methods & Theory: High-dimensional data**

**Wednesday 5 September 4.20pm - 5.20pm**

### **Clustering using Nonparametric Mixtures and Mode Identification**

Shengwei Hu, Yong Wang  
*The University of Auckland*

Clustering aims to partition a set of observations into a proper number of clusters with similar objects allocated to the same group. Current partitioning methods mainly include those based on some measure of distance or probability distribution. Here we propose a mode-based clustering methodology motivated via density estimation and mode identification procedures. The idea is to estimate the data-generating probability distribution using a nonparametric mixture-based density estimator and then locate the modes of the density obtained. In the nonparametric mixture models, each mode and the observations ascend to it correspond to a single cluster. Thus, the problem of determining the number of clusters can be recast as a mode merging problem. A criterion of measuring the importance of a mode is also addressed in this work. The modes would be merged sequentially by its importance until the optimal number of clusters is reached. The proposed method is investigated on both simulated and real-world datasets and achieves sufficiently good performance.

## **7.6 Contributed - Communicating Statistics: Teaching statistics - tips and techniques**

**Wednesday 5 September 4.20pm - 5.20pm**

### **Using learnr to create interactive tutorials for R: from the pinnacle to the pit**

Andy Field

*University of Sussex*

This talk describes the lessons I have learned through developing and teaching with a package of interactive tutorials for R/RStudio (adventr) written using the learnr package (Borges & Allaire, 2017). Part 1: the pinnacle. The first half of the talk highlights the capabilities of learnr and its potential for creating an immersive interactive learning experience through the use of embedded code chunks (with solutions), video and quizzes. I will illustrate these features and discuss how I used learnr in conjunction with flipped classroom teaching to facilitate large group teaching on a postgraduate statistics module. I will reflect upon the benefits of the approach for formative assessment, inclusivity, active and peer-based learning. Part 2: the pit. It didn't all go smoothly. The second half is a journey through various pits of despair into which I unwittingly fell. I will reflect upon the deployment of learnr tutorials (shiny server or a self-contained package?), the potential for learnr tutorials to create a disconnect with R/RStudio, problem areas for teaching data science using learnr (e.g., working with external data files), whether using code solutions hinders learning, and whether the flipped classroom fosters helplessness for less confident students. Throughout the talk I will highlight some areas of potential good (and not so good) practice when using learnr to teach statistics and R/RStudio. I will conclude with suggestions for teaching workflows that aim to maximise the chances of experiencing the pinnacle and not the pit.

## **7.6 Contributed - Communicating Statistics: Teaching statistics - tips and techniques**

**Wednesday 5 September 4.20pm - 5.20pm**

### **The (Q-)Step Change in Demand for Quantitative Social Science: evaluating intervention on a principal-agent problem**

Thomas King

The limited quantitative skills outcomes of undergraduates students of social science in the UK were dubbed market failure and received a substantial intervention to achieve a 'step change'. The ESRC and HEFCE supported a Nuffield Foundation programme, 'Q-Step': funding 15 universities to invest in new and enriched quantitative curriculum. This work analyses the market failure as a principal-agent problem with students reliant on agents in schools, universities and employment to realise efficient market outcomes. Steps are ranged as skills outcomes of awareness, description, research and science. Universities literally transferred curriculum across from more quantitative courses, so stepping up quantitative content in courses and they introduced new, more advanced curriculum at the top end, leading to branded degrees e.g. 'with quantitative methods'. Other activity has improved pedagogy for methods classes and developed the embedding of quantitative inquiry and statistical evidence into substantive teaching. Engagement with schools and employers as outreach, summer schools, placements and dissertation projects have supplemented market signals of course requirements and graduate capabilities. Awareness should be achieved in school, and social description be compulsory, but research skills in design and analysis of empirical study of complex ideas are a step further. These are badly needed in society for occupational analysis activity; more specialised roles require a further science step for commissioning, synthesising, criticising and communicating. Engagement with schools is now supported by the Smith Review recommendations about Core Maths at A level in England which should engender statistically literate awareness. The historically limited engagement with quantification in some programmes and institutions meant a step change could take up general social description. The step change for more advanced courses is more challenging: empirical economics and experimental psychology graduates do not represent ideal outcomes; different social science disciplines use quantification for different abstractions; and the scientific nature of social science has long been neglected for training with tools.

## 7.6 Contributed - Communicating Statistics: Teaching statistics - tips and techniques

Wednesday 5 September 4.20pm - 5.20pm

### Using video feedback when assessing R code

Deirdre Toher

*University of the West of England*

As part of a final year statistics module, students are required to submit a short written report alongside a file containing commented R code that shows the entire of their analysis process. Over the last few years, Deirdre has been using video rather than text-based feedback on this code. She will discuss some of the advantages of using video based feedback on R code submitted by students; which is particularly suited when students are solving open-ended problems where multiple approaches are valid. In such instances, algorithmic assessment is difficult. It also allows emphasis to be put on appropriate commenting of code, giving feedback in a conversational manner. She will also cover some of the lessons learned by experimentation on the best way to ensure that students engage with these videos, including whether or not to release this part of the feedback before, alongside or after releasing the marks to students. She will also raise some of the issues needed to be considered before undertaking to deliver feedback in this manner, including considering how accessible the format is to those with disabilities and to ensuring that students do not have free access to one another's feedback.

## 7.7 Contributed - Data Science

Wednesday 5 September 4.20pm - 5.20pm

### Wind power forecasting using Gaussian process dynamical models

Tobias Jung

*Uniper*

Accurately forecasting the power generated by a wind farm at day-ahead and intraday timescales is the key problem owners and operators of wind farms face and necessary for trading, dispatch, but potentially also other, innovative applications such as offshore maintenance scheduling. Forecasts of wind power are commonly created as either point or probabilistic forecasts on a per look-ahead time basis, i.e., for each lead time of interest separately. Because they model each lead time independently, they are not able to capture temporal interdependence and thus do not properly describe how prediction errors will persist and evolve in time. While this may be sufficient for applications involving single-stage decision making (e.g., trading), it is less appropriate for any kind of application that involves time-dependent or multi-stage decision making (such as in optimal operation of conventional generation in the presence of wind power or optimal planning of distributed storage devices). Various methods have been proposed to this end in the literature, such as time series (Taylor et al., 2009), Gaussian copulas (Pinson et al., 2009), or stochastic differential equations (Moller et al., 2016). In this talk we will investigate Gaussian process state space models (GP-SSM) as a novel tool. GP-SSM have several features which make them particularly attractive: GP-SSM are Bayesian models for dynamical systems with latent state where both the transition and observation model are represented by GPs. Thus no prior constraints are put on the functional form of the transition and complex nonlinear relationships are possible. GP-SSM perform inference both over hyper parameters and latent states and thus deal with nonlinear system identification and adaptive parameters. Inference in GP-SSM can be done efficiently by particle-based MCMC procedures (Svensson et al., 2016). We illustrate the use of GP-SSM for a single wind farm where standard point forecasts from a benchmark industry method are used as the driving force (exogenous information).

## 7.7 Contributed - Data Science

Wednesday 5 September 4.20pm - 5.20pm

### **Toward the Best Discrepancy Forests: A Systematic Comparison with Random Forests**

Jialin Bi,<sup>1</sup> Liang Guo<sup>1</sup>, Ruodan Lu<sup>2</sup>, Jianya Liu<sup>1</sup>

<sup>1</sup>*Shandong University, Weihai*, <sup>2</sup>*University of Cambridge*

Random Forests (RF) have been widely regarded as one of the most effective machine learning algorithms and have received great attentions both in academia and in industry. While RF can achieve remarkable prediction performances by averaging out bias, reducing variance and avoiding overfits, it comes with costs. Ensemble methods usually require many trees and high level of computational resources, making the application of RF cumbersome. Under the circumstance of big data, runtime issue becomes severe due to the scale of data size and the complexity of models. Therefore, there is a pressing need to develop a new algorithm that can retain the advantages of ensemble methods while being fast to train and to deploy. RF's computational burden stems from its resampling mechanism, which relies on a simple random subsampling with replacement for every tree. This mechanism prevents instances from being representatively partitioned to subsets and as a result, more trees are needed to circumvent the risk of obtaining largely dissimilar subsets. We proposed an improved forests classifier method that is based on the "low discrepancy" theory in the field of number theory. In a nutshell, we replaced the random sequence by the best discrepancy sequence to conduct subsampling. We compared the generalization performance of our Best Discrepancy Forests (BDF) with that of RF with 60 publicly available UCI datasets using 50 repeated holdout validations for every dataset. The results showed that BDF significantly outperformed RF in terms of running time. To reach RF's highest level of accuracy, BDF mobilises much less (on average -54.22%) trees and its variance over 50 repetitions of holdout validation is smaller (on average -16.77%) than that of RF. What is more, BDF's highest accuracy score is also 1.56% higher than that of RF. In short, BDF is more accurate and stable while computationally efficient, making it more appropriate for the task of big data analytics.

## 7.7 Contributed - Data Science

Wednesday 5 September 4.20pm - 5.20pm

### **The Influences of Dataset Meta-Features and of Classifier Properties on The Performance of Cross Validation Techniques**

Liang Guo,<sup>1</sup> Jialin Bi<sup>1</sup>, Ruodan Lu<sup>2</sup>, Jianya Liu<sup>1</sup>, Yuqian Cheng<sup>1</sup>

<sup>1</sup>Shandong University at Weihai, <sup>2</sup>University of Cambridge

K-fold-Monte Carlo Cross Validation (MCCV) is a widely-used standard technique for statistical machine learning model and feature selections. However, relying on a random sequence for subsampling pose risks, because some dataset properties prevent MCCV from constructing representative subsets. In addition, the stochastic nature of MCCV makes cross validation results vary a lot from one repetition to another. Finally, MCCV requires significant computational resources, as it must be repeated around 50 times to average out subsampling bias and to converge to the true generalization error values. Therefore, these inherent shortcomings make the generalization performance estimated by MCCV is not an effective one. We argue that MCCV should not be used blindly without taking dataset meta-features and the nature of classifier properties into account. We generated 3,000 of synthetic datasets with various levels of meta-features (i.e. # instances, # features, aspect ratio, modality, # classes, class imbalance, ratio of nominal/numerical features, sparsity, random accuracy, standard deviation ratio, average correlation index, generalized variance, class & feature entropy, mutual information). Each dataset was estimated with the 12 Penn Machine Learning Benchmark classifiers --Gaussian Naïve Bayes(NB), Bernoulli NB, Multinomial NB, Logistic regression, Linear classifier trained via SGD, Linear classifier with passive aggressive algorithm, SVC, KNN, Decision tree, Random forests, AdaBoost, Gradient tree boosting) and with MCCV, Holdout (70%:30%), 0.632 Bootstrap and the K-fold-Best-Discrepancy CV (BDCV, a CV method derived from the field of number theory, which replace MCCV's random sequence with Kuipers & Niederreiter's the best-discrepancy sequence). Results showed that MCCV performed poorly in the datasets with relatively small size, large aspect ratio and imbalance classes. Under most circumstances, the BDCV is the most effective CV in terms of accuracy, variance and running time. We reconducted the experiments using 100 real-world datasets and our conclusions were also held. Our work dismisses the myths and promotes a better understanding regarding to cross validation and generalization performance.

## **7.8 Contributed - Industry & Finance: Dynamic modelling with applications in energy systems**

**Wednesday 5 September 4.20pm - 5.20pm**

### **Dynamic dependence modelling for financial time series**

Georgios Aivaliotis, Yali Dou  
*University of Leeds*

We explore dependence modeling of financial assets in a dynamic way and its critical role in measuring risk. In this paper, we propose two new methods to model the dependence structure of financial assets dynamically: Accelerated Moving Window method and Bottom-up method. The performance of these methods together with the existing Binary Segmentation and Moving Window method is assessed on simulated data. Accelerated Moving Window has the advantage of being applicable in real time and outperforms the standard Moving Window method. This way one could monitor the change in dependence of financial assets and help to warn about wrong model use with implications to the calculation of risk measures. Bottom-up method can only be used to retrospectively fit a dynamic copula model (similarly to Binary Segmentation) and is proved to be the best the method to detect the change of copula although the appropriate minimum sample size can be a problem. The best-performing method is applied to Standard & Poor 500 and Nasdaq indices. Value-at-risk and Expected shortfall are computed from the dynamic copula and the static one respectively to illustrate the effectiveness of dynamic modelling through backtesting.

## **7.8 Contributed - Industry & Finance: Dynamic modelling with applications in energy systems**

**Wednesday 5 September 4.20pm - 5.20pm**

### **Allocation of oil and gas and how to win the lottery**

Phillip Stockton

*Accord Energy Solutions Ltd*

In the North Sea oil and gas industry, hydrocarbons produced offshore from different sources (e.g. platforms) are often mixed together in shared pipelines as they are transported onshore for further processing. In order to identify who owns the hydrocarbons exiting from the pipeline, equations are required to allocate those hydrocarbons at a component level, e.g. propane, butanes, etc. to the sources they were produced from offshore. The talk: Describes and discusses pipeline allocation systems using simple examples to illustrate the fairly simple mathematics involved. It focuses on systems that allocate the discharged products at the end of the pipeline (e.g. entry to an onshore terminal) in proportion to the contributions from each of the upstream entry points (e.g. platform export) as though there was instantaneous transfer. There is no account taken of the transfer time across the pipeline but it is assumed (intuitively) that any differences in the exported versus allocated inlet for each entry point will even themselves out over time. The presentation illustrates how this assumption may not be correct and therefore the allocation is biased, just due to the form of the allocation equations. This is demonstrated using mathematical expectation calculations which are explained simply using the returns that might be expected by buying lottery tickets. Analytical expectation calculations are presented along with the results from Monte Carlo simulations. Though principally explained using simplified examples, real data is also presented which shows the actual impact of these subtle mathematical bias effects resulting in the mis-allocation of millions of pounds worth of hydrocarbons.

## **7.8 Contributed - Industry & Finance: Dynamic modelling with applications in energy systems**

**Wednesday 5 September 4.20pm - 5.20pm**

### **Beyond Optimization via Statistical Emulation and Uncertainty Quantification**

Hailiang Du<sup>1</sup>, Wei Sun<sup>2</sup>

<sup>1</sup>*Durham University*, <sup>2</sup>*University of Edinburgh*

Computer simulators are widely used to make inferences about complex physical systems such as energy systems, in conjunction with historic observations. In energy systems modelling, optimization methods based on certain objective function(s) are widely used to provide deterministic solutions to decision makers. For complex high-dimensional systems, however, simplifications are inevitable to conduct traditional optimization, which leads to the “optimal” solution being suboptimal or nonoptimal. Whilst the optimization problem is well resolved, it would still be valuable for both operational and long term planning purpose to introduce some flexibility to the solution. A novel statistical methodology is introduced, where statistical emulation and uncertainty quantification are employed to identify candidate solutions and to quantify uncertainties (due to i) observational error; ii) emulation approximation; iii) model discrepancy) attach to each candidate solution. Candidate solutions subject to the objective function(s) provide useful flexibility and attached uncertainty quantification provides extra valuable information for decision support. If the model is expensive to evaluate, traditional optimization methods will have serious difficulties, whereas the proposed methodology will not be hampered. It is demonstrated in a practical wind farm planning case study that applying the proposed methodology can overcome challenges in traditional optimization by identifying and inverting the Pareto boundary for decision support.

## 7.9 Contributed - Environmental & Spatial Statistics

Wednesday 5 September 4.20pm - 5.20pm

### Stochastic modelling framework for synthetic time series simulation

Sandhya Patidar

*Heriot-Watt University*

Quite often, observed records (time-series) are not long enough to extract reliable statistics to understand the variability and uncertainty in the patterns and the occurrences of all significant rare events. One possible solution to such issues is to develop robust data-centric modelling techniques for artificially generating realistically possible synthetic sequences based on the historical dataset and within the acceptable statistical errors. The Hidden Markov Model (HMM) is popular stochastic modeling approach and has been applied successfully to model a range of complex processes, such as bio-informatics, speech, molecular evolution, the stock market, natural languages, human and animal behaviour. This presentation is aimed to present the systematic organisation of an HMM-based modeling framework (referred in this presentation and relevant publications as HMM-GP) that couples STL: a Seasonal-Trend decomposition procedure based on Loess and the extreme values distribution (such as Generalised Pareto Distribution), to develop a robust modeling schematic for artificially synthesizing univariate time series. Considering the environmental statisticians as the key audience, the presentation will be organized to demonstrate the step-by-step development of the HMM-GP modeling schematic, including a thorough investigation to examine its suitability for simulating highly stochastic time series of streamflow at a much finer temporal resolution of 15 minutes. A robust validation of the proposed HMM-GP schematic would be presented by conducting an extensive comparison of various statistical characteristics of the observed records with the synthetically simulated flow time series across four hydrologically distinct case-studies River in the UK, namely Don, Nith, Dee, and Tweed. Further, for the benefit of the general audience interested in potential of data-centric modelling techniques and to illustrate the potential of HMM-GP for a wider applicability and transferability across different themes, some key highlights covering the application of HMM-GP in the area of synthetic energy demand synthesis (one/five minutely highly stochastic annual energy demand time series) will also be presented.

## 7.9 Contributed - Environmental & Spatial Statistics

Wednesday 5 September 4.20pm - 5.20pm

### **Tree-Based Inference for Regionalization: A Comparative Study of Global Topological Perturbation Methods**

Mark Janikas,<sup>1</sup> Rodrigo Alves<sup>2</sup>, Renato Assunção<sup>3</sup>

<sup>1</sup>ESRI, <sup>2</sup>Departamento de Ciências Sociais Aplicadas, Centro Federal de Educação Tecnológica de Minas Gerais, <sup>3</sup>Departamento de Ciência da Computação, Universidade Federal de Minas Gerais

Regionalization is a constrained optimization problem that aims to create spatially-contiguous groups, also known as regions. Defining compact, data-driven regions is important to understand patterns in spatial phenomena. Similar to any constrained optimization problem, the spatial constraint may hinder convergence to some global minima, resulting in heterogeneous regions that do not reflect underlying spatial patterns in the data. We present a general methodology for defining compact regions rigorously through perturbing spatial constraints via random spanning trees. In addition, we address questions pertaining to number of optimal regions and probability associated with the regionalization result. Spatial data is represented with a connected graph and spanning trees are used to compress dense graph while preserving spatial relationships and pair-wise similarities. We propose to use consensus-based clustering on different regionalizations of spatial data to perform inference on number of regions. We achieve different regionalizations of spatial data through stochastic perturbation of spatial constraints via random spanning trees. The general framework presented can be used to quantify the effect of the spatial constraints in the overall regionalization result and provide a probabilistic metric for the effect of spatial proximity to value similarity. We propose a spectral-decomposition based heuristic to determine the optimal number of regions in the data. We compare several types of stochastic spanning trees used in inference problems such as fuzzy regionalization and determining number of regions. Performance of stochastic spanning trees are juxtaposed against the traditional permutation-based hypothesis testing frequently used in spatial statistics. Inference results for fuzzy regionalization and determining number of regions is presented on the local area personal incomes for United States counties provided by the Bureau of Economic Analysis. Our results show the evolution of demographic regions (based on economic indicators) in United States over time and show the impact of emerging industries on the overall economic demographics.

## 7.9 Contributed - Environmental & Spatial Statistics

Wednesday 5 September 4.20pm - 5.20pm

### Parametric link functions for spatial generalised linear models

Evangelos Evangelou,<sup>1</sup> Vivekananda Roy<sup>2</sup>

<sup>1</sup>*University of Bath*, <sup>2</sup>*Iowa State University*

Spatial generalized linear mixed models have been popular for analysing spatial data observed in a continuous region. These models assume a prescribed link function that relates the underlying spatial random field with the mean response. On the other hand, there are circumstances, such as when the data contain outlying observations, where the use of a prescribed link function can result in a poor fit which can be improved by the use of a parametric link function. In this talk I will present different sensible choices of parametric link functions which possess certain desirable properties. I will discuss estimation of these models, including model selection and weighing, via multiple importance sampling using the R package *geoBayes*.

## **Keynote 5 - Discussion Meeting: Data visualization**

**Wednesday 5 September 5.30pm - 7.30pm**

### **Visualizing spatiotemporal models with virtual reality: from fully immersive environments to applications in stereoscopic view**

Stefano Castruccio<sup>1</sup>, Marc Genton<sup>2</sup>, Ying Sun<sup>2</sup>

<sup>1</sup>*University of Notre Dame, USA*, <sup>2</sup>*King Abdullah University of Science and Technology*

Recent advances in computing hardware and software present an unprecedented opportunity for statisticians who work with data indexed in space and time to visualize, explore and assess the structure of the data and to improve resulting statistical models. We present results of a 3-year collaboration with a team of visualization experts on the use of stereoscopic view and virtual reality (VR) to visualize spatiotemporal data with animations on non-trivial manifolds. We first present our experience with fully immersive VR with motion tracking devices that enable users to explore global three-dimensional time–temperature fields on a spherical shell interactively. We then introduce a suite of applications with VR mode, freely available for smartphones, to port a visualization experience to any interested people. We also discuss recent work with head-mounted devices such as a VR headset with motion tracking sensors.

## **Keynote 5 - Discussion Meeting: Data visualization**

**Wednesday 5 September 5.30pm - 7.30pm**

### **Visualization in Bayesian workflow**

Jonah Gabry,<sup>1</sup> Daniel Simpson<sup>2</sup>, Aki Vehtari<sup>3</sup>, Michael Betancourt<sup>4</sup>, Andrew Gelman<sup>1</sup>  
*<sup>1</sup>Columbia University, New York, <sup>2</sup>University of Toronto, <sup>3</sup>Aalto University, <sup>4</sup>Columbia University, New York and Symplectomorphic, New York*

Bayesian data analysis is about more than just computing a posterior distribution, and Bayesian visualization is about more than trace plots of Markov chains. Practical Bayesian data analysis, like all data analysis, is an iterative process of model building, inference, model checking and evaluation, and model expansion. Visualization is helpful in each of these stages of the Bayesian workflow and it is indispensable when drawing inferences from the types of modern, high dimensional models that are used by applied researchers.

## **Keynote 5 - Discussion Meeting: Data visualization**

**Wednesday 5 September 5.30pm - 7.30pm**

### **Graphics for uncertainty**

Adrian Bowman

*University of Glasgow*

Graphical methods such as colour shading and animation, which are widely available, can be very effective in communicating uncertainty. In particular, the idea of a 'density strip' provides a conceptually simple representation of a distribution and this is explored in a variety of settings, including a comparison of means, regression and models for contingency tables. Animation is also a very useful device for exploring uncertainty and this is explored particularly in the context of flexible models, expressed in curves and surfaces whose structure is of particular interest. Animation can further provide a helpful mechanism for exploring data in several dimensions. This is explored in the simple but very important setting of spatiotemporal data.

## 8.1 Contributed - Medical: Causal Inference

Thursday 6 September 8.50am - 9.50am

### Benefits and Risks of Radiotherapy in Early Breast Cancer

Sarah Darby,<sup>1</sup> Paul McGale<sup>1</sup>, John Broggio<sup>2</sup>, Jackie Charman<sup>2</sup>, Paul Pharaoh<sup>3</sup>, Gordon Wishart<sup>4</sup>, Jem Rashbass<sup>2</sup>, Carolyn Taylor<sup>1</sup>, Gurdeep Mannu<sup>1</sup>, David Cutter<sup>1</sup>

<sup>1</sup>University of Oxford, <sup>2</sup>Public Health England, <sup>3</sup>University of Cambridge, <sup>4</sup>Anglia Ruskin University

**Background:** Decision aids are widely used to predict the benefits of systemic treatments in early breast cancer, but do not usually assess the effects of radiotherapy. Radiotherapy provides benefit by reducing breast cancer mortality. However it carries risks by increasing mortality from heart disease and lung cancer. Absolute benefit and absolute risk both vary substantially between different women. Nevertheless, they can both be estimated and then combined to predict a woman's likely net benefit from radiotherapy. These calculations are difficult to do in a clinical setting, so radiotherapy may currently be given to some women where predicted risk outweighs predicted benefit. Conversely, it may be withheld from some women where predicted benefit outweighs predicted risk.

**Methods:** Depersonalised individual data from Public Health England on all 600,000 women registered with breast cancer since 1997 were collated. Mortality rates from breast cancer, heart disease, lung cancer and other causes for women who received various treatment combinations were estimated using Poisson regression for women with various characteristics in 5-year categories of time since radiotherapy. These rates were divided by death rate ratios from randomised trials and epidemiological studies to estimate mortality rates in untreated but otherwise similar women. The difference between the two estimates gives the absolute treatment effect. This avoids biases introduced by patient selection for treatment eg due to comorbidities.

**Results:** Predicted 20-year absolute reduction in breast cancer mortality from radiotherapy varied between <0.5% and >6%. Predicted 20-year absolute increase in heart disease and lung cancer mortality from radiotherapy varied between <0.5% and ~5%. The net effect of radiotherapy on 20-year overall survival varied between an absolute gain of ~5% and an absolute loss of ~3%.

**Conclusions:** These estimates, based on large-scale up-to-date population-based data, can be used in decision aids to estimate the likely net effect of radiotherapy for individual women.

## 8.1 Contributed - Medical: Causal Inference

Thursday 6 September 8.50am - 9.50am

### Counterfactual outcome state transition parameters: A new approach to effect heterogeneity for binary outcomes

Anders Huitfeldt

*London School of Economics*

VanderWeele (2012) provided two separate definitions of effect heterogeneity, which he referred to as "effect modification in distribution" and "effect modification in measure". The standard epidemiological approach, which is based on effect modification in measure, is associated with a number of well-described shortcomings, and no consensus exists about whether investigators should define effect heterogeneity in terms of additive or multiplicative measures of effect. More recently, Bareinboim and Pearl (2015) introduced a new graphical framework for transportability, based on effect heterogeneity in distribution. These graphs are an elegant solution to many of the problems associated with traditional approaches, but they require the investigator to make strong assumptions about the data generating mechanism. In light of these limitations, we propose a new definition of effect heterogeneity, based on "counterfactual outcome state transition parameters", that is, the proportion of those individuals who would not have been a case by the end of follow-up if untreated, who would have responded to treatment by becoming a case; and the proportion of those individuals who would have become a case by the end of follow-up if untreated who would have responded to treatment by not becoming a case. Effects are said to be equal between populations if and only if these proportions are equal between the populations. Although counterfactual outcome state transition parameters are generally not identified from the data without strong monotonicity assumptions, we show that when they stay constant between populations, there are important implications for model specification, meta-analysis, and research generalization

*References: Tyler J VanderWeele. Confounding and Effect Modification: Distribution and Measure. Epidemiologic Methods, 1(1):55–82, 8 2012. Elias Bareinboim and Judea Pearl. A General Algorithm for Deciding Transportability of Experimental Results. Journal of Causal Inference, 1(1):107–134, 1 2013.*

## 8.1 Contributed - Medical: Causal Inference

Thursday 6 September 8.50am - 9.50am

### Currently available diagnostics can be unreliable at identifying errors in propensity score specification

Emily Granger, Jamie Sergeant, Mark Lunt  
*The University of Manchester*

In medical research, estimating effects of exposures on outcomes using observational data is challenging due to the inevitability of confounders. Propensity scores provide a potential solution; they can reduce confounder bias by balancing covariate distributions between exposure groups. Despite the recent increase in their use, there is still scepticism about their trustworthiness. One concern is that a misspecified propensity score may not achieve sufficient balance, leading to biased results. There are a variety of diagnostics being used to assess covariate balance after propensity-adjustment, however no consensus on the best way to do this. A simulation study was conducted to compare diagnostics in terms of their ability to identify different types of propensity score misspecification. Diagnostics included are categorised as follows: 1) mean-based, 2) distribution-based or 3) prevalence-based. Categories 1 and 2 respectively include diagnostics which compare covariate means and distributions between treatment groups. Category 3 diagnostics are new; they involve comparing the number of exposed subjects at each covariate value to that predicted by the propensity score. Logistic regression and c-statistics were used to assess how well diagnostics could predict when misspecification occurred. Results indicated that mean-based diagnostics can fail to identify when a nonlinear term is incorrectly omitted from the propensity score model and distribution-based diagnostics are unreliable at identifying omission of interaction terms when sample sizes are small. Prevalence-based diagnostics performed consistently well across all misspecification types and sample sizes and in most cases the regressions failed to converge due to perfect prediction. To help overcome scepticism about propensity scores, being able to reliably assess covariate balance obtained after propensity-adjustment is essential. Unfortunately, our results demonstrate that some of the most widely used (category 1) or readily available (categories 1 and 2) diagnostics can be misleading. Introduction of prevalence-based diagnostics into the applied literature is recommended.

## **8.10 RSS Prize Winners - Best presentations of YSM 2018**

**Thursday 6 September 8.50am - 9.50am**

### **Modelling the UK University Admissions Process**

Ella Kaye, Julia Brettschneider, Anastasia Papavasiliou  
*University of Warwick*

The UK university admissions process is filled with uncertainty for both the applicants and the universities, with both parties needing to make several inter-related decisions throughout the journey. Much of that uncertainty is related to the fact that applications and offers are made on the basis of predicted A-level grades, which is problematic for both parties: over-estimated predications can lead to students accepting offers for which they will not meet the grade conditions, and for universities it can make it hard to ensure the correct cohort size. Moreover, universities want to accept the 'best' students amongst the applicants. We consider six years' worth of admissions data to undergraduate statistics courses at the University of Warwick to explore the accuracy of predicted A-level grades. Along with corresponding first year exam results for those admitted, we explore what in the application might explain success at university. Methods used include logistic regression and proportional odds logistic regression.

## 8.10 RSS Prize Winners - Best presentations of YSM 2018

Thursday 6 September 8.50am - 9.50am

### Why we need randomised trials to study radiotherapy dose in cancer treatment

Johanna Ramroth, Carolyn Taylor, John Broggio, Jackie Charman, Rebecca Elleray, Rebecca Girdler, Geoff Higgins, Gurdeep Mannu, Jason Poole, Michael Skwarski, Ann Watters, Sarah Darby  
*University of Oxford*

**BACKGROUND.** In the era of Big Data, there is increasing interest in using observational data to study radiotherapy dose. We aimed to determine whether observational data could correctly estimate the causal effects on lung cancer survival of increasing radiotherapy dose.

**METHODS.** Non-small-cell lung cancer patients treated with curative intent 2004-2011 were identified in a Thames Valley dataset. Information on potential confounders and outcomes was obtained from other Public Health England sources. Multivariable Cox regressions were conducted.

**RESULTS.** 324 patients were studied. Increasing radiotherapy dose was associated with improved survival in some treatment centres, while in others the opposite was true. These opposite trends suggest that differences in patient selection were driving results, as confirmed in treatment protocols.

**CONCLUSION.** Observational data are not well-suited to estimate causal effects of increasing radiotherapy dose. Long-term differences in effect between doses are likely to be small, and complex patient selection factors overwhelm these effects. Randomised trials thus continue to be the main study design to estimate effects of radiotherapy dose.

*FUNDING.* Financial support for this study was provided by Cancer Research UK (programme grant C8225/A21133 and DPhil studentship OCRC-DPhil12-JR).

## 8.10 RSS Prize Winners - Best presentations of YSM 2018

Thursday 6 September 8.50am - 9.50am

### PSA testing in the UK, is its use justifiable?

Grace Young<sup>1</sup>, Sean Harrison<sup>2</sup>, Emma Turner<sup>2</sup>, Eleanor Walsh<sup>2</sup>, Steven Oliver<sup>2</sup>, Yoav Ben-Shlomo<sup>2</sup>, Simon Evans<sup>2</sup>, Athene Lane<sup>2</sup>, David Neal<sup>3</sup>, Freddie Hamdy<sup>3</sup>

<sup>1</sup>Population Health Sciences, <sup>2</sup>University of Bristol, <sup>3</sup>University of Oxford

**Background:** Two major trials have assessed screening with prostate-specific antigen (PSA) for prostate cancer with conflicting results on mortality. A new larger trial based in the UK (the CAP trial) was designed to investigate this further.

**Methods:** The CAP trial randomised practices to invite men for a single PSA test (screening) or usual practice and recruited ~400,000 men between 2001 and 2009. The primary endpoint was prostate cancer mortality rate at a median 10 years of follow up. We conducted an additional retrospective cohort study on 450,000 men from CPRD aged 45-69 to examine the risk of receiving a PSA test in the UK between 2002 and 2012.

**Results:** The men in CPRD had a 39.2% risk (95% CI 39.0% to 39.4%) of receiving a PSA test over the 10-year period. These results help contextualise the recent findings of the CAP trial, which found that there was no evidence for a difference in prostate cancer mortality when comparing men invited for a screening versus usual practice; RR 0.96 (95% CI 0.85 to 1.08).

**Discussion:** The risk of receiving a PSA test in the UK is very high, despite there being no formal screening programme in place and no proof of benefit.

## 8.2 Contributed - Official Statistics & Public Policy: Using Data for Public Good

Thursday 6 September 8.50am - 9.50am

### How integrated data can be used for public good

Becky Tinsley

*Office for National Statistics*

The Office for National Statistics has been using integrated data to research into the Government ambition that “censuses after 2021 will be based on alternative sources of data.” The Administrative Data Census project has made progress in: producing a range of Research Outputs that are typically produced by the ten-yearly census; comparing these outputs with official statistics; and seeking feedback from users. Research so far has covered the size of the population, household statistics (including the size and structure of households), and a range of population characteristics including income, labour market status, commuting flow patterns, ethnicity and most recently, nationality. These outputs have used a range of integrating methods and data sources including administrative data, commercial (aggregate mobile phone data) and survey data. We are now expanding our research to look beyond what is traditionally produced by the census to understand how integrated data can be used to provide new insights into society. Using integrated data brings a range of new challenges in the production of statistics to meet user needs such as: understanding the data quality and definition differences; developing new methods to measure and adjust for coverage errors; and producing multivariate small area outputs when variables come from different data sources with different coverage errors. A key part of making the best use of the vast range of data available is to demonstrate the public good of using these data. This involves explaining: Why the statistics are important to shaping public policy The benefits these statistics bring to the public How we are protecting the data we are using This presentation will explore the new types of analysis that is possible from integrated data, the social benefits of producing these outputs and our plans to promote understanding of these benefits.

## **8.2 Contributed - Official Statistics & Public Policy: Using Data for Public Good**

**Thursday 6 September 8.50am - 9.50am**

### **Measuring National Well-being: Insights into the differences in quality of life and well-being across age groups**

Rhian Jones, Silvia Manclossi  
*Office for National Statistics (ONS),*

In November 2010, the Measuring National Well-being (MNW) Programme was established. The aim was to monitor and report “how the UK as a whole is doing” by producing accepted and trusted measures of the well-being of the nation. This ongoing work is part of an initiative, both in the UK and internationally, to look beyond traditional headline economic growth figures to establish a fuller picture of social progress. The first task of the MNW Programme involved a National Debate conducted in 2010/11 in which people across the UK were asked ‘what matters most’ to them. Based on their responses, an indicator set was created covering ten areas of life including our health, the natural environment, personal finances and crime. Up to that point, policy makers had tended to focus their decisions on maximising economic growth. However, this programme has helped decision makers look at the world around them differently. Once a year, we report progress against this set of headline indicators in a publication called, *Life in the UK*. We now also provide a visual overview of the data through the MNW online dashboard, which can be explored either by specific areas of life or by the direction of change (i.e. whether areas have improved or deteriorated). In a better effort to understand how different groups experience life in the UK and where inequalities exist, for this year publication, we have specifically analysed the measures of MNW where data by age is collected for younger and older people. This presentation is based on the most recently available data as of April 2018 with a particular focus on findings about quality of life and well-being across different age groups, drawing on the results of the latest publication of *Life in the UK 2018*.

## 8.2 Contributed - Official Statistics & Public Policy: Using Data for Public Good

Thursday 6 September 8.50am - 9.50am

### **Collaboration across government and beyond to enable better decisions to be made about the handling of domestic abuse**

Alexa Bradley

*Office for National Statistics*

Domestic abuse occurs in many forms, and although often hidden, it is widespread. It shatters the lives of victims and their families. But it isn't handled adequately across the criminal justice system. Her Majesty's Inspectorate of Constabulary recommended data should be brought together to provide better understanding of how domestic abuse is handled, to improve victims' experiences of criminal justice and encourage more victims to come forward. ONS led the production of a bulletin and data tool to address this and, in 2017, expanded the coverage of data sources to provide a more comprehensive and coherent picture. We brought voluntary sector data together with government statistics for the first time to fill an identified gap – insight into cases that don't enter the criminal justice system but are supported by the voluntary sector. Individual data sources don't provide the context needed to understand the full picture and our publication provides more transparency and coherence, giving valuable insight into an important area of public policy. By bringing different data sources together we were able to show that many victims do not see justice. The majority of cases do not come to the attention of the police, and many of those that do come to their attention do not result in a conviction for the perpetrator. We were also able to highlight variations in the provision of services for victims across areas, and that whilst other agencies such as health and social care are already involved in the response to domestic abuse, such involvement is not widespread. Our products are actively helping organisations to identify areas for improvement and make more informed decisions about how they can help victims. They have provided valuable evidence that underpins the government proposals on transforming the response to domestic abuse. With our better information, improved service delivery and decision making is helping change lives for the better.

### 8.3 Contributed - Applications of Statistics: Prisons and Pensions

Thursday 6 September 8.50am - 9.50am

#### Statistical significance - meaningful or not

Colin Aitken,<sup>1</sup> Amy Wilson<sup>1</sup>, Richard Sleeman<sup>2</sup>

<sup>1</sup>*The University of Edinburgh*, <sup>2</sup>*MSA Ltd., Filton, Bristol*

Statistical tests with large sample sizes can have large power. Power is the ability to detect an effect. Detection is indicated by a result which is statistically significant. A test with large power will detect a very small effect. This very small effect may not be meaningful in the context of the analysis being conducted. The courts have the perception that for an effect to be meaningful it is necessary for the effect to be statistically significant. However, statistical significance is not a sufficient condition for an effect to be meaningful. This can lead to a difficulty where testimony of no meaningful effect is interpreted by counsel as one of no statistically significant effect. Should the difference between a meaningful effect and a statistically significant effect be explained in reports and if so, how? Some possible answers are proposed.

### 8.3 Contributed - Applications of Statistics: Prisons and Pensions

Thursday 6 September 8.50am - 9.50am

#### **What works in the rehabilitation of people who have offended? A meta-analysis of JDL findings**

Sandra von Paris, Joanna Adler, Mark Coulson  
*Middlesex University*

The Ministry of Justice Data Lab (JDL) offers data analysis services with reference to UK reoffending administrative data to organisations working with people who have offended, to evaluate the effectiveness of rehabilitation interventions based on the reduction of various re-offending measures. JDL reports show promising effects on re-offending measures in terms of magnitude, but often without statistical significance. Meta-analyses were performed drawing together effect sizes of individual interventions to analyse them in context with each other and to assess the statistical significance of the summary effect. Various moderating effects relating to cohort, intervention, and organization characteristics were investigated to identify factors shaping outcomes, including intervention type, sector, and region. The spectrum of moderator data types dictated the required meta-analysis model types: fixed, random and mixed effects meta-regression models were employed. Data limitations in relation to model-overfitting were investigated by increasing model complexity and dimensionality through combination of moderators. To evaluate model quality across multiple moderators of different data types and combinations thereof, various visual representation and summary methods were developed. This allowed the comparison of subgroup meta-analyses and meta-regressions with continuous moderators to find the best models. The moderators defining the best models may be considered as the factors best explaining differences between intervention outcomes. Overall effects size estimates for re-offending measures were statistically significant and implied overall positive intervention impacts. Some significant moderators were identified in relation to the characteristics of the intervention participants. This indicates that intervention group characteristics not only influence their future re-offending behaviours but also the effects an intervention might have on them and how these emerge. As more data becomes available from further JDL studies, the methodological search for significant moderators amongst available cohort, intervention, and organization characteristics may yield important information on how best to design interventions to help people who have offended minimise re-offending behaviours and maximise re-integration into society.

## 8.3 Contributed - Applications of Statistics: Prisons and Pensions

Thursday 6 September 8.50am - 9.50am

### How to value pension funds

Jane Hutton

*The University of Warwick*

Applied statisticians enjoy inter-disciplinary discussions. Pension provision is changing, as beliefs about life expectancy and returns on investments can lead to large estimated deficits. This has resulted in several pension funds, such as Royal Mail, moving from defined benefit to defined contribution schemes. It is natural to ask how reliable and robust the estimates are. What happens when a medical statistician thinks from first principles about approaches to estimating the liabilities of a pension fund and the assets and income required to meet those liabilities? The liabilities depend on demographic variables such as life expectancy, age at retirement, marriage rates, ill-health retirement and death in service rates. The Pension Regulator expects an evidence-based approach to estimating the future demographic profile of scheme members. Large funds can obtain good estimates from members' data, small funds will need to consider national data. The United Kingdom benefits from a long history of national statistics, whereas countries such as Tanzania are less fortunate. Estimation of death rates in Tanzania has to rely more on sampling and interpolation, and is therefore less precise. Approaches to estimating life tables taken by demographers, actuaries and medical statisticians differ, as do methods or criteria for good estimates and propagation of uncertainty. Various factors which affect life expectancy which are common knowledge among epidemiologists might not be familiar to actuaries. For example, the 'widowhood effect' is well documented, but seems not to be considered in estimating joint life expectancies in pension funds. Medical journals typically require clear statements of the data used, and of the uncertainty in the results presented. Statisticians can help to improve estimation of liabilities and assets, and the presentation of the information to pension fund members. One might even consider the use of probabilistic decision analysis.

## **8.4 Contributed - Social Statistics:**

**Thursday 6 September 8.50am - 9.50am**

### **Counting Welsh speakers in Wales: Measuring reality not perception**

Eilir Jones

*Freelance*

Determining the number of Welsh speakers is of crucial importance in a wide range of policy and social areas in Wales. This is especially the case given the recently set official Welsh Government target of one million speakers by 2050, almost double the current number. The question is an ever-increasing challenge for social statisticians. First language speakers are becoming fewer and there is an increasing number of second-language speakers spanning a wide range of fluency levels, making it more difficult to agree how to ask someone to classify themselves as a Welsh speaker. The main source of information on number of speakers is the census question, which is simple and granular and provides decennial trend data. However, it doesn't ask about fluency (only ability to read, write and/or speak). It also permits one person answer on behalf of the rest of the household, the main impact of which that children's Welsh language ability is grossly overstated. There's also no information on each person's use of Welsh, nor whether it's the household language (as asked for all languages other than Welsh and English), nor any figures on Welsh speakers in the rest of the UK. I will outline various options for evaluating the number of Welsh speakers and their use of Welsh, so that policy-makers and the public can have confidence in the figures. Among the issues addressed are the attainment of a simple framework for gauging fluency, the challenge of gauging fluency among children, the errors associated with different existing measures and the viability of a UK-wide count. I will also give examples from broadcasting, where assessing fluency is crucial given its direct impact on measuring viewing and listening to Welsh language content. Croeso i Gymru! Welcome to Wales and to share in our challenge.

## 8.4 Contributed - Social Statistics:

Thursday 6 September 8.50am - 9.50am

### Validating non-participation bias methodology using register-based health surveys

Megan Yates<sup>1</sup>, Pekka Martakainen<sup>2</sup>, Tommi Härkänen<sup>3</sup>, Oarabile Molaodi<sup>1</sup>, Hanna Tolonen<sup>3</sup>, Alastair Leyland<sup>1</sup>, Lindsay Gray<sup>1</sup>

<sup>1</sup>MRC/CSO Social and Public Health Sciences Unit, University of Glasgow, <sup>2</sup>Department of Sociology, University of Helsinki, Finland, <sup>3</sup>National Institute for Health and Welfare (THL), Finland

Population-sampled studies provide surveillance of health and related determinants, yet decreasing participation levels threaten the validity of estimates purported to be representative of their target population. The standard means of correcting for such bias include weighting the participant's response data, usually based upon sociodemographic characteristics. However, typically this does not adequately correct for differences between participants and the target population within the sociodemographic sub-groups. We aim to explore the validity of recently developed alternative methodology, with application to refine the measurement of alcohol consumption. In this new method, record-linkage between survey data and administrative hospitalisation and death information at the individual level is used with reference to data on the general population to infer on non-participants. The inference is used as the basis of generating partial synthetic observations for non-participants including the corresponding rates of hospitalisations and deaths due to alcohol related causes in demographic sub-groups; multiple imputation is then applied to fill-in their "missing" alcohol measurements for combining with observations for participants in order to obtain refined estimates of population consumption. In order to validate the methodology, we make use of the Health 2000 survey conducted in Finland, and an 11% sample of the contemporaneous population, with follow-up until 2012. As Finland maintains a population-wide, individual-level register of all residents, the age, sex, socioeconomic measures (education and socioeconomic status), and alcohol-related hospitalisation and death records are available for both the participants and – crucially (although their alcohol consumption remains unknown) – non-participants. As a basis for the test of the methodology, comparison of the population and participant samples will be used to generate synthetic data on non-participants. Alcohol consumption for both the actual and synthetic data on non-participants will then be multiply imputed and overall alcohol consumption estimates can be compared between the two approaches to evaluate the performance of the methodology.

## 8.4 Contributed - Social Statistics:

Thursday 6 September 8.50am - 9.50am

### **Transform or translate? How to optimise uptake of on-line social surveys.**

Ian O'Sullivan, Laura Wilson, Andrew Phelps, Alex Nolan  
ONS

The ONS Strategy Better Statistics, Better Decisions aims to transform how business and household data is sourced. A key feature of the strategy is the use of non-survey data as the primary and default data source for future statistical outputs. Direct data collection via surveys will feature in this alternative model but to a lesser extent than at present, and when surveys are required they will be mixed-mode. Moving voluntary social survey data collection on-line will bring challenges, but likewise will provide survey methodologists with the opportunity to take a different approach to designing questionnaires. A previous change programme, the Electronic Data Collection Transformation Programme tried to 'simply' translate (lift and shift) the current Labour Force Survey into an on-line mode. This approach, which aligns with a more traditional data-user centred approach left the survey content equally as long, confusing and repetitive as some of our existing surveys, and would certainly not make for a good 'user experience'. Likewise this 'translate' approach comes with a risk of poor on-line uptake and poor quality data. As part of the Census and Survey Transformation Programme, ONS are now taking a more contemporary 'user-centric' approach, which simply means putting the respondent in the driving seat when it comes to the design of the survey experience. Government Digital Service (GDS) principles are being embraced as part of the redesign of social surveys. Incidentally 'user-centric' is not a new term, having driven the development of products in the Tech World, with the aim of creating something which has high 'usability' for decades. This presentation will provide further details of the user-centric approach being taken to redesign the LFS, as well as reporting on the quantitative evidence from several large-scale trials of the online questionnaire.

## 8.5 Contributed - Methods & Theory: Regression and ranking models

Thursday 6 September 8.50am - 9.50am

### Issues in Modelling Rankings Data

Heather Turner,<sup>1</sup> Jacob Van Etten<sup>2</sup>, David Firth<sup>3</sup>, Ioannis Kosmidis<sup>3</sup>

<sup>1</sup>*Freelance*, <sup>2</sup>*Bioversity International*, <sup>3</sup>*University of Warwick and The Alan Turing Institute*

The Plackett-Luce model is a well-established approach for modelling rankings data such as the finishing orders in a set of races or consumer preferences from market research. However there can be many issues with applying this model in practice: observed rankings often have features not accommodated by the model, and the available software can be inefficient, without the facility for model-based inference. This presentation describes how these issues are addressed by methodology implemented in the recently released R package, `PlackettLuce`. We propose a novel generalization of the Plackett-Luce model, which can accommodate ties in the rankings, as well as partial rankings (rankings of only a subset of items). To ensure the parameters always have finite maximum likelihood estimates and standard errors, we use pseudo-rankings: additional wins and losses between each item and a ghost item. This makes it possible to handle clustered rankings (rankings for distinct sets of items) or rankings in which one or more items are always ranked first or last in their rankings. We estimate the parameters of our model using iterative scaling or direct maximization of the likelihood (for example with the Broyden–Fletcher–Goldfarb–Shanno algorithm). This approach does not require expanding the rankings into individual choices or construction of large model matrices with dummy variables. Therefore it scales well to a moderate number of items and a large number of unique rankings. With an efficient implementation of a single model fit, we are able to use model-based partitioning to fit Plackett-Luce trees that allow for heterogeneity in item worth, for example due to different judges making rankings. Aspects of our approach will be illustrated via the motivating application of a citizen science project in agricultural development.

## 8.5 Contributed - Methods & Theory: Regression and ranking models

Thursday 6 September 8.50am - 9.50am

### Leverage-weighted least squares estimation

Keith Knight

*University of Toronto*

In linear models, the idea of "bounded influence" estimation dates back more than 40 years with the work of Colin Mallows among others. In this talk, we will consider some simple methods for controlling the influence of an arbitrary subset of  $k$  observations. Ordinary and weighted least squares estimates in linear regression models can be expressed as weighted means of so-called elemental estimates, which are estimates based on subsets of  $p$  observations where  $p$  is the dimension of the vector of regression parameters. The weight for each elemental estimate is simply the determinant of a  $p \times p$  submatrix (whose row and column indices correspond to the observations defining the elemental estimate) of a projection matrix (which, in the case of ordinary least squares, is the "hat" matrix). Using this formulation, we can then define a leverage score for a subset of  $k$  observations for any weighted least squares problem. This leads us to consider, for example, weighted least squares estimates where we maximize the entropy of the weights subject to a constraint on the leverage score. Alternatively, we can define estimates that lie on a parameterized path whose endpoints are ordinary least squares estimates and the "leave-out- $k$ " least squares estimates.

## 8.5 Contributed - Methods & Theory: Regression and ranking models

Thursday 6 September 8.50am - 9.50am

### Time-varying Regression Slope Estimation for Time series data

Sucharita Ghosh

*Statistics Lab, Swiss Federal Research Institute WSL, Birmensdorf, Switzerland*

Two global temperature series containing mean annual anomalies from land and sea surfaces are plotted against each other suggesting a possible regression-type relationship. This is one of many examples, where estimation of regression slopes is of interest using time series data. It is well-known that in such cases, both time series must be carefully detrended before a regression analysis is carried out. Also, presence of auto-correlations in the data can cause spurious trend-like patterns affecting precision of statistical estimates. Consider for instance the two-stage regression model with smooth but arbitrary trends and a constant slope  $\beta$  : (1)  $E[ Y_i | X_i ] = \alpha(t_i) + \beta X_i$  , and (2)  $E[X_i] = \theta(t_i)$ , where,  $t_i=i/n$ ,  $i=1,2,\dots,n$ . In this case, in spite of the time-varying functions  $\alpha(t)$  and  $\theta(t)$  in the model, constant slope estimation can easily be done via the usual method of least squares using regression residuals instead of the raw  $(X_i, Y_i)$  data. In contrast, however, if the slope parameter is not a constant but an arbitrary smooth function of time  $\beta(t)$ , the usual least squares approach for slope estimation fails. In this talk, we describe a kernel based procedure when the errors have auto-correlations as well as non-Gaussian marginal distributions. For instance, transformations of latent Gaussian processes can lead to such errors. We discuss optimal bandwidth (window-width) selection strategies and numerical examples. For the global temperature series for instance, which spans more than 150 years, it turns out that allowing the regression slope to be time-dependent reveals some interesting patterns in the data, which however remains hidden if just a constant slope model is used. (Source of global temperature series: MET Office, UK).

## **8.6 Contributed - Communicating Statistics:**

**Thursday 6 September 8.50am - 9.50am**

### **What can we learn about the teaching of statistics from the historical relationship between sociology and statistics in Britain?**

Plamena Panayotova

*The University of Edinburgh*

What can we learn about the teaching of statistics from the historical relationship between sociology and statistics in Britain? This talk focusses on the teaching of statistical methods and thinking to social scientists by drawing on a research on the historical relationship between statistics and sociology in Britain. An analysis of the teaching of sociology and statistics at the London School of Economics and Political Science (LSE) in the twentieth century and of sociology methods syllabuses from the majority of universities in the UK in 1968 and 1979 suggest that, except in the early decades of the twentieth century, statistical methods have regularly been present in the sociology curriculum. However, there is also substantial historical evidence suggesting that the potential of the available statistics teaching has consistently been undermined by sociologists' understanding of sociology as aiming to provide general education; their preoccupation with theoretical knowledge and neglectful attitude towards the place of methodology in sociology teaching; and by an anti-quantitative culture nurtured on the basis of attitudes and lack of substantial statistical knowledge. Drawing on numerous examples from archival sources, the ultimate aim of this talk is to show what we can learn from history in order to improve the teaching of statistics to social scientists. Teaching simply more statistics is not going to be enough. Until statistics is perceived as an 'esoteric activity' and communicated narrowly as formulas, commands and tests that should be mastered, social scientists will continue to miss the point about the meaning of statistics as a generally applicable thinking tool and as a worldview. The question is even more relevant now, since not only adequate and intelligent study of society, but active democratic participation in society in the twenty-first century, is becoming increasingly difficult without statistical literacy.

## **8.6 Contributed - Communicating Statistics:**

**Thursday 6 September 8.50am - 9.50am**

### **Online, interactive data skills modules - making learning more accessible**

Vanessa Higgins, Sarah King-Hele  
*UK Data Service, University of Manchester*

This presentation will showcase the new online, interactive 'Data Skills Modules' that have been created by data specialists within the UK Data Service in collaboration with an e-learning technologist. Containing a mix of videos, quizzes, written materials and activities, these learning resources aim to get users started with survey, longitudinal or aggregate data. Each of the three modules is approximately two-hours long and introduces a different data type. The modules are designed for learners who want to get to grips with data, highlighting key features, showing how to use the data and include short videos and interactive quizzes to test the learner's knowledge. The Survey Data Module includes sections about looking at characteristics using graphs and looking at two variables to understand patterns, the Longitudinal Data Module extends this to look at methods for exploring longitudinal data and the Aggregate Data Module includes a section about mapping census data. The Data Skills Modules are freely available for everyone without registration and are designed to be conducted in the learner's own time, dipping in and out when needed. The modules are free to access and open access software is also used where possible. The presentation will also include reflections and experiences of producing the modules and how the modules have been received by the community. The modules are on the UK Data Service website: <https://www.ukdataservice.ac.uk/use-data/data-skills-modules>

## 8.6 Contributed - Communicating Statistics:

Thursday 6 September 8.50am - 9.50am

### Language and images that support the conceptual understanding of Statistical Inference

Hilary Watt

*Imperial College*

There is widespread misunderstanding of statistical inference; the logic and language used can feel counter-intuitive. We can change the core language that we use in teaching inference, to align our language more closely with our practical applications. I suggest sparse use of terms such as “point estimate”. I prefer the more concrete phrase “results obtained in study participants”. This may avoid the errors in interpreting non-significant results as “no difference”, even when differences are found in study participants. We can ground the definition of confidence intervals in our desire to apply knowledge to some greater population. Confidence intervals are “interval estimates, based on the random sampling variation inherent in generalising beyond study participants, to the population from which they were selected at random”. We can discuss their practical value and interpretation when our participants are not a random sample of any population. The frequent confusion of standard errors with standard deviations, may be reduced by describing standard errors as measures of precision. They measure the precision, based on random sampling variation, that results from using our participant effect size to estimate the population effect size. A diagram can show correspondence between the ratio of the effect size to their standard errors (z-values) and p-values (with written interpretations). This may result in greater understanding than the common practice of showing the standardised Normal curve. However, I acknowledge that such Normal distributions aid understanding of the mechanism of calculation of p-values. In this context, the Normal distribution corresponds to the distribution of sample means, a challenging concept. Student understanding may be supported by pointing out that “the sampling distribution of means” is an artificial distribution. It arises only when developing Statistical theory. It differs from the distribution of repeated estimates obtained in practical research, because real examples come from different sample sizes, and from slightly different settings (and hence from different populations).

## 8.7 Contributed - Medical: Linear Models and Multiple Testing

Thursday 6 September 8.50am - 9.50am

### Epidemiology and information theory: R2, entropy and information

Deborah Schneider

*Imperial College London*

In epidemiology and medical research, there is a strong interest in assessing how well a set of risk factors explain the variation of an outcome. Classically, researchers look for exogenous factors (known or unknown) and fit a model around them for one or several outcomes of interest. In this context, we are concerned with goodness-of-fit. This is well understood in linear models, where we can use the R<sup>2</sup> coefficient of an information criterion. With the development of large throughput sequencing technologies and availability of -omics data, there is an increased interest in non-linear, more complex models. In particular, -omics data have (very) high dimensions and require different analysis approaches. Similarly, -omics data tends to be used in highly non-linear models, for which classical methods fail to capture the complex nature of the association between the exogenous and endogenous variables. In order to address this, works have been done using the Shannon entropy and the mutual for the purpose of defining “generalised R<sup>2</sup>” measures (Hastie, T., 1987) or “coefficients of determination” (Eshima & Tabata, 2010), which aim to extend to models evolving in the exponential family and beyond linearity. While the generalisation of these methods is yet to be investigated, they also have estimation issues, as they rely on the KL measure. In this paper, we review these two methods and assess how well they generalise – or even apply for linear models. Additionally, we propose an extension that builds on the Fisher information – a quantity analogous to entropy – which has similar asymptotic properties and is more computationally tractable. We apply this new method together with the two aforementioned to the modelling of Systolic blood pressure and cardiovascular incidents in UKBiobank.

## 8.7 Contributed - Medical: Linear Models and Multiple Testing

Thursday 6 September 8.50am - 9.50am

### **Multiple testing approaches for evaluating the effectiveness of a drug combination in a multiple-dose factorial design.**

Saswati Saha

*University of Bremen*

Drug combination trials are often motivated from the fact that producing a completely new drug is very expensive and using existing drugs in combination might prove to be more fruitful and less expensive than using the component drug alone. Several approaches have been explored for developing statistical methods that compare (single) fixed dose combination therapies to its component. But extension of these approaches to a multiple dose combination clinical trial is not always so simple. We have proposed three approaches by which one can provide confirmatory assurance that combination of two drugs is more effective than either component drug alone. These approaches involved multiple comparisons in multilevel factorial design where the overall type 1 error is controlled by bonferroni test, bootstrap test and lastly by considering the least favorable null configuration under a union intersection test. We have thus built a R package implementing the above approaches and in this presentation we would like to demonstrate how this can be used in a drug combination trial. Further we want to demonstrate using extensive simulations, how these three approaches are performing when bench marked with an existing approach.

## 8.7 Contributed - Medical: Linear Models and Multiple Testing

Thursday 6 September 8.50am - 9.50am

### **Robust mixed modelling: A new approach to handle time-varying outlier impacts**

Laura Boyle,<sup>1</sup> Lisa McFetridge<sup>1</sup>, Özgür Asar<sup>2</sup>

<sup>1</sup>Queen's University Belfast, <sup>2</sup>Acıbadem Mehmet Ali Aydınlar University, Istanbul

Medical research increasingly involves the analysis of longitudinal data, where multiple observations are made on a set of variables for each patient over time. Linear mixed effects models are frequently utilised to model the influence of both fixed effects, such as age and treatment plan, and random effects, representing latent individual characteristics, on a longitudinal response. It is common for longitudinal outliers to occur in medical research, both in scenarios where the observations for a specific patient outlie from the population (b-outliers), and where observations made on an individual patient contain outlying values (e-outliers). Previous research has shown that longitudinal outliers violate the normality assumptions of standard mixed effects models, thereby introducing inefficiency into the parameter estimation process. Both e-outliers and b-outliers can be accommodated by using robust mixed models, in which the standard normality assumptions are replaced with t-distributional assumptions. However, there is currently no methodology which can account for the impact of time-varying outlier patterns in longitudinal data. It is commonly witnessed in various medical applications that patients take time to respond to new treatments, frequently experiencing a period of time where their response is more likely to fluctuate and produce observations which outlie from the expected. Time-varying outlier methodology would capture such scenarios and thus be invaluable in medical research. This type of time-varying outlier pattern is observed within a dataset collected by the Northern Ireland Renal Information Service between 2002 and 2012, containing information on 1320 haemodialysis patients with a total of 27,113 repeated measurements. This research presents a novel time-varying outlier impacts (TOI) methodology for robust mixed models, which enables the robustness parameters to evolve over time. Results from the utilisation of TOI mixed models will be presented from application to the aforementioned renal data on haemodialysis patients, in addition to simulated data.

## 8.8 Contributed - Methods & Theory: Missing and censored data

Thursday 6 September 8.50am - 9.50am

### Bayesian missing data models with weights

Harvey Goldstein,<sup>1</sup> James Carpenter<sup>2</sup>

<sup>1</sup>*University of Bristol*, <sup>2</sup>*London School of Hygiene*

Many datasets, especially from surveys, are made available to users with weights. Where the derivation of such weights is known, this information can often be incorporated into the user's substantive model (model of interest). When the derivation is unknown, the established procedure is to carry out a weighted analysis. However, with non-trivial proportions of missing data this is inefficient and may be biased when data are not missing at random. Bayesian approaches provide a natural approach for the imputation of missing data, but it is unclear how to handle the weights. In this presentation, we propose a weighted bootstrap MCMC algorithm for estimation and inference. This extends existing work by allowing interaction and polynomial terms to be included in the model of interest as well as allowing the fitting of complex multilevel models. A simulation study shows that the procedure has good inferential properties. We illustrate its utility with an analysis of data from the Millennium Cohort Study.

## 8.8 Contributed - Methods & Theory: Missing and censored data

Thursday 6 September 8.50am - 9.50am

### A non-proportional hazards MPR model for interval censored data.

Gilbert MacKenzie<sup>1</sup>, Defen Peng<sup>2</sup>, Kevin Burke<sup>1</sup>

<sup>1</sup>University of Limerick, <sup>2</sup>BC Centre for Improved Cardiovascular Health, University of British Columbia

The Weibull multi-parameter regression (MPR) model with Gamma frailty [1] and structural dispersion [2] is developed for interval censored survival data. The basic MPR model which models the scale and shape parameters simultaneously by means of two linear predictors is wholly parametric with non-proportional hazards. It was developed by Burke & MacKenzie in their seminal 2017 Biometrics paper [3]. We describe the basic model, develop the interval-censored likelihood and extend the model to include Gamma frailty and structural dispersion. In addition, we present a simulation study and re-analyse data from the Signal Tandmobiel dental study in relation to gender and dmf [5]. An analysis of the non-parametric maximum likelihood estimators of the cumulative hazard functions (for boys and girls) shows that the time-to-emergence of teeth<sup>24</sup> [6] follows a power curve which provides empirical support for the MPR Weibull model. The structural dispersion component suggests some difference in frailty between the sexes and this is confirmed when fitting separate models for boys and girls, the frailty variance being larger in boys. Overall, the MPR models provide a better fit to the data than do their single parameter (PH and non-PH) competitors. MPR models are relatively new and are of increasing interest to statisticians working on survival analysis. Historically the idea of modelling the scale and shape parameters in the hazard function symmetrically has been under-appreciated.

Key References Hougaard, P. (2000) Analysis of multivariate survival data, Springer-Verlag, New York. Lynch, J and MacKenzie, G. (2014). Frailty models with structural dispersion. In: Statistical Modelling in Biostatistics and Bioinformatics, MacKenzie & Peng, Springer Verlag. (ISBN 978-3-319-04578-8). Burke, K. and MacKenzie, G. (2017). Burke K. and MacKenzie G. (2017). Multi-Parameter Regression Survival Models – an alternative to Cox model. Biometrics Volume 73, Issue 2, pp 678–686.

## **8.8 Contributed - Methods & Theory: Missing and censored data**

**Thursday 6 September 8.50am - 9.50am**

### **On the use of the Not at Random Fully Conditional Specification Procedure (NARFCS) in Practice**

Daniel Tompsett

*MRC Biostatistics Unit, University Of Cambridge*

In this talk we will describe the Not at Random Fully Conditional Specification (NARFCS) procedure, a formal adaptation of the Fully Conditional Specification (FCS) procedure to impute data under Missing not at Random (MNAR) conditions. In particular we note that the procedures sensitivity parameters, which determine the extent to which imputed data depart from the Missing at Random (MAR) assumption, condition on all other variables in the imputation process. This makes them very difficult to elicit from expert opinion. We show that misunderstanding the conditional nature of these sensitivity parameters will often lead to imputed data's that depart from MAR in ways inconsistent with the users assumptions. We therefore describe a means to calibrate the procedures sensitivity parameters so that the resultant imputed data's show differences between observed and missing individuals consistent with simpler sensitivity parameters, which are not conditional on other variables in the data, which we can elicit instead. We also discuss the possibility of the inclusion of the missingness indicators of the data variables as part of the imputation process, reasoning that they may themselves predict missingness in other variables. The procedure is demonstrated on a dataset from the Avon Longitudinal Study of Parents and Children (ALSPAC)

## 8.9 Contributed - Environmental & Spatial Statistics

Thursday 6 September 8.50am - 9.50am

### **A spatial regression model for the disaggregation of areal unit based data to high resolution grids with application to vaccination coverage mapping**

Chigozie Utazi<sup>1</sup>, Julia Thorley<sup>1</sup>, Victor Alegana<sup>1</sup>, Matthew Ferrari<sup>2</sup>, Andy Tatem<sup>1</sup>

<sup>1</sup>University of Southampton, <sup>2</sup>Pennsylvania State University

A spatial regression model for the disaggregation of areal unit based data to high resolution grids with application to vaccination coverage mapping Utazi CE1, Thorley J1, Alegana VA1, Ferrari M2 and Tatem AJ11 University of Southampton, UK and 2 Pennsylvania State University, USA Abstract This paper develops a methodology for high resolution mapping of vaccination coverage using areal data in settings where geolocated survey data are inaccessible. The proposed methodology is a binomial spatial regression model with a logit link and a combination of covariate data and random effects modelling two levels of spatial autocorrelation in the linear predictor. The principal aspect of the model is the melding of the misaligned areal data and the prediction grid points using the regression component and each of the conditional autoregressive and the Gaussian spatial process random effects. The Bayesian model is fitted using the INLA-SPDE approach. We demonstrate the predictive ability of the model using simulated data sets depicting real-life settings. The results obtained indicate a good predictive performance by the model. The methodology is applied to predicting the coverage of measles and diphtheria-tetanus-pertussis vaccinations at 1x1 km in Afghanistan and Pakistan using subnational Demographic and Health Surveys data. The predicted maps are used to highlight 'coldspots' of low vaccination coverage and assess progress towards vaccination targets to facilitate the implementation of more geographically precise interventions.

## 8.9 Contributed - Environmental & Spatial Statistics

Thursday 6 September 8.50am - 9.50am

### Sampling and oversampling units with prescribed characteristics: an adaptive design proposal

Federico Andreis<sup>1</sup>, Marco Bonetti<sup>2</sup>

<sup>1</sup>University of Stirling, <sup>2</sup>Bocconi University, Milan, Italy

Investigating a rare and geographically clustered trait in a finite population is challenging: traditional designs require large sample sizes to obtain accurate estimates, resulting in a considerable investment of resources. Moreover, over-representation of units with prescribed characteristics may also be a desirable feature, especially when resources are limited: budget constraints usually limit the survey effort, hence the need for an efficient use of resources. Notable examples include: i) WHO's tuberculosis (TB) prevalence surveys, crucial for countries that bear a high TB burden (still less than 1%): in the latest guidelines, spatial patterns are not explicitly accounted for, with the risk of missing a large number of cases, and ii) expensive environmental field studies needed to identify local maxima such as sources of pollution, areas where soil erosion is reaching critical levels, or where individuals of rare species are observed: the ability to obtain samples where the sought-after characteristics are more likely to appear, is then very important. Adaptive designs are typically well suited to surveying populations where the variable of interest has a highly skewed distribution, also in a geographical sense; for example, when dealing with a dichotomous survey variable such as the presence or absence of a rare and geographically clustered characteristic, adaptive strategies have proved to be reliable in over-representing units that have the trait of interest, while retaining the possibility of drawing valid inference. We introduce a novel method to extract a sample from a finite population where units with desired characteristics are over-represented. The approach is both sequential and adaptive and allows, via suitable compositions of predictive and objective functions, to target specific subsets of the population. We consider the problem of design-based estimation and conjecture the validity of a modified Horvitz-Thompson estimator capable to account for the imbalance induced by the targeting procedure. We demonstrate the potential of our proposal via simulation.

## **Keynote 6 – Barnett Lecture**

**Thursday 6 September 10.00am - 11.00am**

### **Analyse problems, not data**

Peter Diggle

*Lancaster University and Health Data Research UK*

A major challenge for the statistics discipline in the twenty-first century is to respond positively to the rise of data science in order to maintain the relevance of statistical method (singular) as an integral component of scientific method. As the twenty-first century progresses, effective statistical science will increasingly depend on close collaboration between statisticians and subject-matter experts, with the analysis strategy reflecting not only the data-format, but also the purpose of the analysis, i.e. a shift in focus from data to problems.

In this talk, I shall offer a personal perspective on this challenge, in the specific context of spatial and spatio-temporal modelling of public health data.

## **9.1 Medical: STREngthening Analytical Thinking for Observational Studies (The STRATOS Initiative)**

**Thursday 6 September 11.30am - 12.50pm**

### **Review and comparison of spline procedures in multivariable regression.**

Aris Perperoglou

*University of Essex*

Splines offer high flexibility for modelling complex variable forms for continuous covariates within a regression setting. This flexibility requires the user to have a good understanding of how to select an appropriate spline function and how to tune parameters to obtain an optimal fit. However, in practice, many researchers will use software for spline fitting at the provided default values. This approach of using off-the-shelf software often leads to errors and highlights the need for guidance in the use of splines. In a recent paper members of the Topic Group 2 of the STRATOS Initiative (<http://www.stratos-initiative.org/>), reviewed R packages that include functions for spline modelling within a regression framework. In this talk we will illustrate challenges that an analyst faces when working with data and showcase differences in spline fits that can be attributed to the choice of hyper-parameters rather than the basis used. We will compare between methodological approaches and provide practical guidance on available software for building multivariable Gaussian, logistic and Cox regression models using splines. We will also use simulated datasets to investigate how spline methods can be used in a multivariable setting where selection of variables is of interest.

## **9.1 Medical: STREngthening Analytical Thinking for Observational Studies (The STRATOS Initiative)**

**Thursday 6 September 11.30am - 12.50pm**

### **The STRATOS initiative, illustrated by issues in Topic group 2: selection of variables and their functional forms**

Willi Sauerbrei for the STRATOS initiative

*Medical Center University of Freiburg*

Research questions have become more complex, stimulating continuous efforts to develop new and even more complex statistical methods. Tremendous progress in methodology has been made, but has it reached researchers who analyze observational studies? Part of the underlying problem may be that even experts do often not agree on potential advantages and disadvantages of competing approaches. However, many analysts are required de facto to make important modelling decisions and would be delighted to receive guidance. In most observational studies selection of variables and determination of the functional form for continuous variables is required. There is general agreement that subject matter knowledge should play a key role to determine suitable models but is there sufficient knowledge to determine a model without important data-dependent decisions? What would constitute a 'state-of-the-art' analysis? Many variable selection strategies have been proposed, and for dealing with continuous variables at least four strategies are used. Most analysts choose either (1) step functions (based on categorization) (2) assume that the functional form is linear (3) use fractional polynomials or (4) use one of the many approaches based on spline functions. While (1) and (2) have clear disadvantages and should not be the final step in the analysis, (3) and (4) have been criticized as well, for different reasons. The STREngthening Analytical Thinking for Observational Studies (STRATOS) Initiative (<http://www.stratos-initiative.org/>; Sauerbrei et al, *Statistics in Medicine*, 2014, 33: 5413-5432) is an international collaboration of researchers which was formed to help bridge the gap between methodological innovation and application by developing guidance for researchers. Discussing key issues of Topic group 2 'Selection of variables and their functional forms in multivariable analysis' we will illustrate the concept, structure and the general approach. It will become obvious that considerable research is required to gain better insight into advantages and disadvantages of competing strategies.

## 9.1 Medical: STRENGTHENING ANALYTICAL THINKING FOR OBSERVATIONAL STUDIES (THE STRATOS INITIATIVE)

Thursday 6 September 11.30am - 12.50pm

### How to impute missing data in Cox regression: New developments incorporating non-proportional hazards

Ruth Keogh<sup>1</sup>, Tim Morris<sup>2</sup>

<sup>1</sup>London School of Hygiene and Tropical Medicine, <sup>2</sup>MRC Clinical Trials Unit at UCL

Missing data is a problem in many time-to-event analyses and it is well known that the 'complete-case' approach of dropping individuals with missing data can result in bias and loss of efficiency. White and Royston (Statistics in Medicine 2009) and Bartlett et al. (SMMR 2015) described two different multiple imputation (MI) methods suitable for use with Cox regression. Part of the model building process in Cox regression analyses is to test the proportional hazards assumption and potentially to estimate time-varying effects of exposures. However, no MI methods have been devised which handle time-varying effects of exposures. I will describe extensions of the two MI methods to this setting and show that these perform well. It will also be shown that ignoring time-varying effects at the imputation stage results in incorrect tests for proportional hazards, biased estimates of time-varying effects and substantial loss of power in detecting time-varying effects. I will focus on time-varying effects modelled using restricted cubic splines and will outline a model building strategy that incorporates both MI and selection of time-varying effects. I will present some results from simulation and real-world examples. R code is available for applying the proposed methods (<https://github.com/ruthkeogh/MI-TVE>).

## **9.2 Official Statistics & Public Policy: Post-Brexit and Post-Wales Act – what statistics do we need in Wales?**

**Thursday 6 September 11.30am - 12.50pm**

### **Post-Brexit and Post-Wales Act – what statistics do we need in Wales?**

James Tucker,<sup>1</sup> Glyn Jones<sup>2</sup>

<sup>1</sup>*Office for National Statistics*, <sup>2</sup>*Welsh Government*

We are entering a post-European Union world and, separately, more powers are being devolved to Wales including the first Welsh taxes in 800 years. Do we have the statistics we need to be able to develop public policy – social, economic and cultural – in this new constitutional landscape? This interactive panel discussion is presented by the RSS South Wales Local Group, and brings together experts from business, financial services, academia and government to discuss key questions such as How can official statistics and the academic community meet these challenges together? What role can new data sources and techniques have in helping us understand Wales? There are statistical and conceptual issues about developing statistics at a sub-national or local level. How can we ensure such data are fit for purpose? Confirmed panellists include representatives from Welsh Government, Development Bank Wales and the Institute of Welsh Affairs.

### 9.3 Applications of Statistics: Emerging applications in statistics

Thursday 6 September 11.30am - 12.50pm

#### **Outcome-driven mixture modelling of longitudinal data with side information for precision medicine : study of cognitive decline**

Anaïs Rouanet, Rob Johnson, Sylvia Richardson, Brian Tom  
*MRC Biostatistics Unit*

Precision medicine is based on the analysis of extensive information to better characterise a disease and to identify subgroups of patients with distinct mechanisms of disease or particular responses to treatments. According to a patient characteristics shared with a subgroup of similar individuals, clinicians can then use these research findings to diagnose accurately and differentially, monitor the disease course, predict the risk and the response to treatments at the individual level. New statistical methodology developments focus on the integration of rich biological, clinical and epidemiological data now available to get to ever-finer patient stratification. The objective of this work was to develop an outcome-driven mixture model to uncover biologically and clinically meaningful subgroups of the population, associated with specific cognitive evolution patterns and brain imaging profiles. Profile regression is a Bayesian model-based clustering method, linking non-parametrically an outcome and side information (additional attributes providing partitioning information) through cluster membership. A notable advantage of this method is that the number of clusters is unconstrained. Johnson et al. (2010) extended it to a longitudinal outcome using a Gaussian Process, and we considered different kernels for the covariance function to fit best epidemiological data. We applied the model on a sample from the North American Alzheimer's Disease Neuroimaging Initiative cohort. We focused on the Mini-Mental State Examination cognitive test, collected during 8 years, and on volumetric MRI brain imaging biomarkers. We identified 4 clusters, with a clear distinction between the imaging profiles of the cognitively stable clusters and the ones with low or declining patterns, giving insight into the heterogeneity in cognitive decline. This methodology offers a flexible clustering tool, handling a longitudinal outcome and side information. Such a model can facilitate the early identification of subjects at high risk of cognitive decline, providing a useful tool for clinical decision-making in a precision medicine framework or for patient recruitment in clinical trials.

### 9.3 Applications of Statistics: Emerging applications in statistics

Thursday 6 September 11.30am - 12.50pm

#### **A Spatial Modeling Approach for Linguistic Object Data: Analysing dialect sound variations across Great Britain**

Shahin Tavakoli,<sup>1</sup> Davide Pigoli<sup>2</sup>, John Aston<sup>3</sup>, John Coleman<sup>4</sup>

<sup>1</sup>*University of Warwick*, <sup>2</sup>*King's College London*, <sup>3</sup>*University of Cambridge*, <sup>4</sup>*University of Oxford*

Dialect variation is of considerable interest in linguistics and other social sciences. However, traditionally it has been studied using proxies (transcriptions) rather than acoustic recordings directly. We introduce novel statistical techniques to analyse geolocalised speech recordings and to explore the spatial variation of pronunciations continuously over the region of interest, as opposed to traditional isoglosses, which provide a discrete partition of the region. Data of this type require an explicit modeling of the variation in the mean and the covariance. Usual Euclidean metrics are not appropriate, and we therefore introduce the concept of d-covariance, which allows consistent estimation both in space and at individual locations. We then propose spatial smoothing for these objects which accounts for the possibly non convex geometry of the domain of interest. We apply the proposed method to data from the spoken part of the British National Corpus, deposited at the British Library, London, and we produce maps of the dialect variation over Great Britain. In addition, the methods allow for acoustic reconstruction across the domain of interest, allowing researchers to listen to the statistical analysis. This is joint work with Davide Pigoli (King's College London), John Aston (Cambridge), and John Coleman (Oxford).

### 9.3 Applications of Statistics: Emerging applications in statistics

Thursday 6 September 11.30am - 12.50pm

#### The Endless Quest for Understanding Terrorism

Andre Python,<sup>1</sup> Janine Illian<sup>2</sup>, Charlotte Jones-Todd<sup>2</sup>, Marta Blangiardo<sup>3</sup>

<sup>1</sup>University of Oxford, <sup>2</sup>University of St Andrews, <sup>3</sup>Imperial College London

Terrorism is a global threat to peace and security, as exemplified by the ongoing lethal attacks perpetrated by ISIS in Iraq, Al Qaeda in Yemen, and Boko Haram in Nigeria. Despite almost half a century of research analysing data contained in databases on terrorism, the causes that lead to terrorism are still not well understood. In this talk, I first identify key barriers that have prevented academic researchers to generalise their findings. Second, I provide a brief overview of major scientific work that has contributed to a better understanding of terrorism at different temporal and spatial scales. I conclude my talk by discussing recent studies that illustrate new insights into the nature of terrorism that have been revealed by applying Bayesian geostatistical models to spatio-temporal data.

## **9.4 Social Statistics: Using statistics to understand, predict and evaluate domestic abuse**

**Thursday 6 September 11.30am - 12.50pm**

### **Can we predict domestic abuse?**

Ruth Weir

*University of Essex*

Understanding and reducing domestic abuse has become an issue of priority for both local and national governments in the UK, with its substantial human, social and economic costs. It is a subject that has been researched across many disciplines, but no research in the UK to date has focused on neighbourhood level predictors of domestic abuse and their variation across space. This paper uses Geographically Weighted Regression (GWR) to model structural and cultural predictors of police reported domestic abuse. Readily available structural and characteristic variables were found to predict the domestic abuse rate and the repeat victimisation rate at the Lower Super Output Area level and the model coefficients were all found to be non-stationary, indicating varying relationships across space. Conducting this research not only has important implications for victims' wellbeing, but it also enables policy makers to gain a better understanding of the geography of victimisation, allowing targeted policies to be implemented, resources to be more efficiently allocated and their impact evaluated.

## **9.4 Social Statistics: Using statistics to understand, predict and evaluate domestic abuse**

**Thursday 6 September 11.30am - 12.50pm**

### **Simple Statistics and Domestic Abuse**

Matthew Bland

*Cambridge Centre for Evidence Based Policing*

Domestic abuse has come to the forefront of public attention in the last two decades. Today it generally accounts for more than 10% of all the crimes reported to police forces, and impacts on other public agencies from education to health services, private business and third-sector bodies. The level of financial and specialist resource investment in dealing with domestic abuse is appropriate for such a major public health concern, yet still relatively little is known about this very private form of crime. Police databases offer one of the richest sources of potential knowledge, but until recently have remained separated and unexplored. In this talk, Matthew Bland will present some of the findings from his research into police data using mostly simple, descriptive statistics to establish a common set of facts and take on the challenge of confirming or refuting common practitioner beliefs. The big questions this talk tackles include: - Just how much domestic abuse is repeated?- How many 'serial' perpetrators are there out there?- Does domestic abuse always get more serious over time?- How harmful is domestic abuse?The talk concludes with a discussion of how practitioners may take these facts forward to shape services in the future.

## **9.5 Methods & Theory: Advances and recent applications in two-phase designs**

**Thursday 6 September 11.30am - 12.50pm**

### **Semiparametric Inference for Merged Data from Multiple Sources**

Takumi Saegusa

*University of Maryland*

Nowadays, every organization collects massive data from multiple heterogeneous sources. The representativeness of these data sets often depends on technology of data collection but such technology does not ensure equal access to the target population. A potential remedy to reduce selection bias is to merge data with different target populations. The main statistical issue is (1) potential duplicated selection from overlapping sources, (2) unidentified duplication and (3) randomness due to sampling from population and sampling from data sources. The resultant sample is then a biased and dependent sample with duplication. By viewing two-phase stratified sampling as data integration from non-overlapping sources, theory and methods from the analysis of stratified samples continue to prove useful. In this talk, I extend weighted semiparametric estimation and inverse probability weighted empirical process theory for two-phase stratified sampling to data integration problems. We address the necessity of the extension of empirical process theory for the analysis of semiparametric inference by illustrating how empirical process results are used to develop consistency, rate of convergence and asymptotic normality of the weighed estimator. Finite sample properties of the proposed estimator is studied in simulation studies in the Cox proportional hazards model.

## **9.5 Methods & Theory: Advances and recent applications in two-phase designs**

**Thursday 6 September 11.30am - 12.50pm**

### **On the analysis of two-phase designs in cluster-correlated data settings**

Sebastien Haneuse<sup>1</sup>, Claudia Rivera-Rodriguez<sup>2</sup>, Donna Spiegelman<sup>1</sup>

<sup>1</sup>*Harvard T.H. Chan School of Public Health*, <sup>2</sup>*University of Auckland*

In public health research information that is readily available may be insufficient to address the primary question(s) of interest. One cost-efficient way forward, especially in resource-limited settings, is to conduct a two-phase study in which the population is initially stratified, at phase I, by the outcome and/or some categorical risk factor(s). At phase II detailed covariate data is ascertained on a sub-sample within each phase I strata. While analysis methods for two-phase designs are well established, they have focused exclusively on settings in which participants are assumed to be independent. As such, when participants are naturally clustered (e.g. patients within clinics) these methods may yield invalid inference. To address this we develop a novel analysis approach based on inverse-probability weighting (IPW) that permits researchers to specify some working covariance structure, appropriately accounts for the sampling design and ensures valid inference via a robust sandwich estimator. In addition, to enhance statistical efficiency, we propose a calibrated IPW estimator that makes use of information available at phase I but not used in the design. A comprehensive simulation study is conducted to evaluate small-sample operating characteristics, including the impact of using naïve methods that ignore correlation due to clustering, as well as to investigate design considerations. Finally, the methods are illustrated using data from a one-time survey of the national anti-retroviral treatment program in Malawi

## 9.5 Methods & Theory: Advances and recent applications in two-phase designs

Thursday 6 September 11.30am - 12.50pm

### Generalized Meta-Analysis for Multivariate Regression Models Across Studies with Disparate Covariate Information

Nilanjan Chatterjee

*Johns Hopkins University*

Meta-analysis, because of both logistical convenience and statistical efficiency, is widely popular for synthesizing information on common parameters of interest across multiple studies. We propose developing a generalized meta-analysis (GMeta) approach for combining information on multivariate regression parameters across multiple different studies which have varying level 20 of covariate information. Using algebraic relationships between regression parameters in different dimensions, we specify a set of moment equations for estimating parameters of a maximal model through information available from sets of parameter estimates from a series of reduced models available from the different studies. The specification of the equations requires a reference dataset to estimate the joint distribution of the covariates. We propose to solve these 25 equations using the generalized method of moments approach, with the optimal weighting of the equations taking into account uncertainty associated with estimates of the parameters of the reduced models. We describe extensions of the iterated reweighted least square algorithm for fitting generalized linear regression models using the proposed framework. Methods are illustrated using extensive simulation studies and a real data example involving the development of a breast 30 cancer risk prediction model using disparate risk factor information from multiple studies.

## **9.5 Methods & Theory: Advances and recent applications in two-phase designs**

**Thursday 6 September 11.30am - 12.50pm**

### **Multi-wave, Outcome Dependent Sampling Designs for Longitudinal Binary Data**

Jonathan Schildcrout

*Vanderbilt University Medical Center*

Retrospective outcome dependent sampling (ODS) designs are an efficient class of study designs that may be implemented when resource constraints prohibit ascertainment of an expensive covariate on all members of a cohort. One type of ODS design for longitudinal binary data stratifies individuals into three strata according to those who never, sometimes, and always experience the binary event at the observed follow-up times. For time-varying covariate effects, it has been shown that sampling only individuals with response variation (i.e., those who sometimes experience the event) yields highly efficient estimates. If inference lies in a time-invariant covariate effect, or in the joint effect of time-varying and time-invariant covariates, then the design choice is not clear. Since the ideal design for many estimation targets is not always obvious, we propose a class of multi-wave ODS designs for longitudinal binary data where later wave designs are identified after data have been collected and examined at earlier waves. We will describe the class of designs, examine finite sampling operating characteristics, and apply the designs to an exemplar longitudinal cohort study, the Lung Health Study.

## 9.8 Medical: Stepped-Wedge Challenges

Thursday 6 September 11.30am - 12.50pm

### Sample size for stepped wedge trials

Richard Hooper

*Queen Mary University of London*

In a cluster randomised trial it is well understood that failure to allow for clustering in the sample size calculation will result in an underpowered study. Stepped wedge designs add new layers of complexity to sample size calculation. Key to this is an understanding of the multi-level structure of the outcome data. Stepped wedge trials follow the same clusters longitudinally, but depending on the design they may assess each individual just once. Where the adjustment to the sample size for a classic cluster randomised trial is concerned with the intracluster correlation (the correlation between outcomes of two individuals from the same cluster), the adjustment for a stepped wedge trial must consider how this correlation changes depending on when the individuals were sampled. While classic cluster randomised trials are inherently inefficient, a stepped wedge trial design allows comparison of intervention and control within the same cluster, so that a judicious choice of design can actually reduce the sample size needed to achieve given power. The methodological literature has seen an explosion of interest in all aspects of stepped wedge trial design, matching the exponential rise in numbers of published protocols for stepped wedge trials. Trialists can now access a growing collection of software tools and other resources to help them design stepped wedge trials.

## 9.8 Medical: Stepped-Wedge Challenges

Thursday 6 September 11.30am - 12.50pm

### Practical workshop on sample size calculation for efficient cluster and stepped wedge randomised trials

Karla Hemming

*University of Birmingham*

Determining sample size required for cluster randomised trials involves a complex interplay between practical (i.e., cost or logistical) constraints and statistical efficiency considerations (i.e., the desire to maximise statistical power whilst minimising the total sample size). Design of these trials might simply involve determining the power or detectable difference for a given number of clusters and average cluster size. However, with a drive to maximise the social and ethical value of trials, researchers often need to consider the trade-offs between recruiting more clusters versus increasing cluster sizes. Furthermore, when alternative designs are feasible, researchers might be able to reduce the number of clusters or participants per cluster by adopting a more efficient design such as the cluster cross-over trial, the cluster randomised trial with a baseline assessment, or the stepped-wedge design. Determining the sample size or power for these alternative designs has seen a flurry of activity in the methodological literature. Of important note, the usual intra-cluster correlations become more complex when measurements are taken at anything other than a single point in time. In this workshop we introduce and illustrate a web-based tool, called an “R Shiny app” which will allow researchers to implement these methodological advances; and which will also allow researchers to clearly appreciate not only the conventional trade-offs between cluster size and number of clusters; but will also allow them to consider the statistical efficiency trade-offs between different cluster designs; and the impact of uncertainty of key correlation parameters.

## 9.8 Medical: Stepped-Wedge Challenges

Thursday 6 September 11.30am - 12.50pm

### Issues related to the Efficiency of the stepped wedge design

Simon Bond

*Cambridge University Hospitals*

A standard idealised step-wedge design satisfies the requirements, in terms of the structure of the observation units, to be considered a Balanced Design and can be labelled as a criss-cross design (time crossed with cluster) with replication. As such Nelder's theory of general balance can be used to decompose the analysis of variance into independent strata (grand mean, cluster, time, cluster:time, residuals). If time is considered as a fixed effect then the treatment effect of interest is estimated solely within cluster and time:cluster strata; the time effects are estimated solely within the time stratum. This separation leads to directly to scalar, rather than matrix, algebraic manipulations to provide closed-form expressions for standard errors of the treatment effect. We use the tools provided by the theory of general balance to obtain expression for the standard error of the estimated treatment effect in a general case where the assumed covariance structure includes random-effects at the time and time:cluster level. This provides insights that are helpful for experimental design regarding the assumed correlation within clusters over time, sample size in terms of numbers of clusters and replication within cluster, and components of the standard error for estimated treatment effect.

## **9.8 Medical: Stepped-Wedge Challenges**

**Thursday 6 September 11.30am - 12.50pm**

### **Brief introduction to the stepped wedge design**

Mona Kanaan

*University of York*

This talk will give a brief introduction to the stepped wedge randomised controlled trial design, its history, current status and challenges.

## **9.9 Environmental & Spatial Statistics: Spatial inequalities in disease risk: How unequal is society?**

**Thursday 6 September 11.30am - 12.50pm**

### **Hierarchical age-period-cohort models for spatially heterogeneity in mortality trends**

Theresa Smith

*University of Bath*

Age-period-cohort (APC) models have been used to understand incidence and mortality trends for diseases like cancer since the 1980s. However, fitting and interpreting these models requires great care because of a well-known identifiability problem: given any two of age period, and cohort, the third is determined. There is a growing interest in using APC models to investigate discrepancies in incidence or mortality between multiple strata such as sexes, regions, or disease sub-types. Jointly modelling the evolution in time trends of disease across multiple subgroups must be done carefully to avoid further identifiability problems. In this talk I review APC models and an identifiable parametrization fit in a Bayesian framework (as in Smith and Wakefield 2016). I extend this model to allow for spatially correlated sets of age, period and cohort effects, which can be fit using MCMC (e.g., in Stan) or INLA. I conclude with an application to mortality data in the European Union, illustrating how the hierarchical APC models can be used to impute mortality for countries with shorter data series.

## **9.9 Environmental & Spatial Statistics: Spatial inequalities in disease risk: How unequal is society?**

**Thursday 6 September 11.30am - 12.50pm**

### **Extending interrupted time series for matched small area data with spatial discontinuity: an application to Municipal Waste Incinerator effects.**

Anna Freni-Sterrantino  
*Imperial College London*

**Abstract** In an Interrupted Times Series study (ITS), a time series is used to establish an underlying trend, which is 'interrupted' by an intervention at a known point in time. Such methods have been extensively applied in epidemiology, in the evaluation of care quality and to assess the impact of changes in health and public health policies. In its original formulation, ITS considers only the time series of the outcome of interest and does not account for controls; however, the inference may be biased if the outcome exhibits a time trend which might be confounded with the intervention. We propose a framework for applying the interrupted time series approach to small area data to answer the question if the opening of a new Incinerators, acting as an intervention, influences infant mortality and mean birth. We extended the ITS as follows: (i) we defined and model matched spatial controls as buffers similar to the exposed areas for extension and population characteristic; (ii) we specified spatial dependency allowing for spatial discontinuity generated by exposed and control buffers; (iii) we performed joint a Bayesian inference of exposed and control areas which allow uncertainty to be propagated across; and (vi) we defined two synthetic indexes to evaluate the presence of the effect of the intervention and its strength on the two outcomes. We applied our approach on eight Incinerators in England and Wales, of 10km buffers and time series 1996-2012. The approach is suitable for the analysis of quasi-experimental time series studies where additional dependency (e.g., spatial, non-linear time trend) is present and controls are needed.

## **9.9 Environmental & Spatial Statistics: Spatial inequalities in disease risk: How unequal is society?**

**Thursday 6 September 11.30am - 12.50pm**

### **Refining global models of social inequality and ill health with local data**

Dianna Smith<sup>1</sup>, Claire Thompson<sup>2</sup>

<sup>1</sup>*University of Southampton*, <sup>2</sup>*London School of Hygiene and Tropical Medicine*

**Objectives:** The aim of this project is to modify a global model of expected social inequality to better capture local-level population profiles. The starting point is a model to predict food poverty risk in small areas (Middle or Lower Super Output Areas) and the intention is to adapt it to better estimate risk in areas such as rural regions of England.

**Methods:** The initial model estimates risk using population data from the 2011 Census, updated using mid-year population estimates from the Office for National Statistics and Department for Work & Pensions data on benefit claimants to estimate risk of household food poverty in England. External validation is based on a comparison with childhood obesity data from the National Child Measurement Programme. This work has been published, however, we are keen to revise the model to address other factors limiting food security in households. Stakeholders indicate that rural areas are problematic to model as well as areas of high migration. Here we explore options to incorporate new variables to update models for a variety of area typologies.

**Results:** The main challenge to updating models for different areas is adequate data availability and quality. Where this is available from local government there is an opportunity to revise original model specifications to improve estimation.

**Conclusions:** Data availability is crucial to creating area-specific models, however, with continuing collaborative research a template for local data collation from existing sources and an online interface will make this more feasible across England. This may be a valuable tool for local public health and Clinical Commissioning Groups.

## 10.1 Contributed - Medical: Survival Analysis

Thursday 6 September 2pm - 3pm

### **Predicting the chances of remission from epileptic seizures over time**

Laura Bonnett<sup>1</sup>, Jane Hutton<sup>2</sup>, Anthony Marson<sup>1</sup>, Catrin Tudur Smith<sup>1</sup>, David McLernon<sup>3</sup>  
<sup>1</sup>University of Liverpool, <sup>2</sup>University of Warwick, <sup>3</sup>University of Aberdeen

Previous prediction models within epilepsy have generally only been able to predict outcome at a single point in time. Whilst this provides an initial prognosis, it is insufficient for making clinically meaningful treatment choices and informing patient counselling. It is far more useful to assess how the prognosis of each individual changes over time in order to balance the risks and benefits of a variety of treatment options with expectation management. Additionally, epilepsy models are usually based on covariate information available at randomisation within clinical trials, or the start of treatment. In real life the situation is more complicated – people with epilepsy have regular clinic appointments during which they report information including number of seizures since previous appointment. This ever-changing information is directly related to the condition of the patient and can be used for better individualised treatment depending on a dynamic assessment of the prognosis of the patient. A dynamic prediction model was therefore developed which estimated the chances of 12-month remission in people with epilepsy over a six year horizon of opportunity. These predictions were recomputed each month to provide a dynamic assessment of the individualised chances of remission whilst taking account of seizure count since baseline. This is the first time such a model has been applied to epilepsy data. A number of clinical variables were predictive of 12-month remission from seizures. The predictions from our model therefore have the potential to help clinicians better manage the expectations of people with epilepsy not only at the start of treatment, but throughout their epilepsy journey. External validation is now necessary to strengthen the generalisability and transportability of the model.

## 10.1 Contributed - Medical: Survival Analysis

Thursday 6 September 2pm - 3pm

### Survival analysis based on high-dimensional genomic data

Arief Gusnanto,<sup>1</sup> Khaled Alqahtani<sup>2</sup>, Henry Wood<sup>3</sup>, Charles Taylor<sup>1</sup>

<sup>1</sup>University of Leeds, <sup>2</sup>Prince Sattam bin Abdulaziz University, <sup>3</sup>Leeds Institute of Cancer and Pathology

Copy number alterations (CNA) are structural variation in the genome, in which some regions exhibit more or less than the normal two chromosomal copies. This genomic CNA profile provides critical information in tumour progression and is therefore informative for patients' survival. It is currently a statistical challenge to model patients' survival using their genomic CNA profiles while at the same time identify regions in the genome that are associated with patients survival. Some methods have been proposed, including Cox proportional hazard (PH) model with ridge, lasso, or elastic net penalties. However, these methods do not take the general dependencies between genomic regions into account and produce results that are difficult to interpret. In this paper, we propose a new formulation of sparse-smoothed Cox PH model (SSCox) that takes into account general dependencies between genomic regions while simultaneously performing variable selection via sparse solution. Our formulation makes Cox PH model with ridge, lasso, and elastic net penalties as special cases of the SSCox method. The results indicate that the proposed SSCox method shows a better prediction performance than the other models in our comparison, while enabling us to investigate regions in the genome that are associated with the patients' survival with sensible interpretation. We illustrate the method using a real dataset from a lung cancer cohort and simulated data.

## 10.1 Contributed - Medical: Survival Analysis

Thursday 6 September 2pm - 3pm

### How competing risks can bias estimated hazard ratios in survival analysis

John Gregson, Linda Sharples, Jonathan Bartlett, Stuart Pocock,  
*LSHTM, Astra Zeneca*

Background: A competing risk is an event that prevents an event of interest, such as a primary trial outcome, from occurring. The most common competing risk in biostatistics is death.

Objectives: To understand interpretation of hazard ratios when competing risks are present.

Methods: We analysed two case studies using: 1) Proportional hazards (PH) models which censor patient follow-up at a competing event and assume no relationship with the event of primary interest; 2) Fine and Gray models in which patients remain "at risk" following a competing event; 3) PH models with multiple imputation of events that might have occurred had the competing risk not been present; 4) parametric joint frailty models which allow for dependence between the competing event and event of interest.

Results: We estimated dementia risk in obese versus normal weight patients in a large clinical database. Death, a competing event, occurred more frequently in obese patients. Using PH models obesity was associated with decreased dementia risk. With Fine and Gray models, obesity was more strongly associated with decreased dementia risk, reflecting that obese patients had higher mortality rate and therefore developed dementia less frequently. In sensitivity analyses using multiple imputation and joint frailty models, which both assumed positive correlation between death and dementia risk, obesity was less strongly associated with decreased dementia risk than using PH models. Similar results were obtained when these methods were applied to a clinical trial, in which the primary outcome of hospital admission was interrupted by death as a competing event.

Conclusions: PH models do not account for the dependence between competing events and the event of primary interest, potentially leading to biased hazard ratios. Using Fine and Gray models may make matters worse. Sensitivity analyses using multiple imputation or joint frailty help to understand the impact competing risks have on hazard ratios for effect estimates of primary interest.

## 10.2 Contributed - Official Statistics & Public Policy: 50th Anniversary of UK National Housing Surveys

Thursday 6 September 2pm - 3pm

### 50 years of Welsh housing conditions

Jenny Davies,

*Welsh Government (Llywodraeth Cymru)*

The first survey of housing conditions in the UK took place in 1967 and covered England and Wales. In 1968 Wales conducted its first such survey, making 2018 the 50th anniversary of housing condition survey in Wales. The purpose of this presentation is to give a view of how the condition of the Welsh housing stock has improved over the last 50 years and asks, 'what next for housing conditions surveys?' Initially the focus of the surveys was whether homes had basic amenities: a bath or shower; an indoor WC; a wash hand basin; and hot and cold water at three points. At that time 26% of dwellings were without an inside WC! Over time, as conditions have improved, the focus has shifted more to energy performance/efficiency and the social context of housing. The survey methodology has also changed. For example, the survey instrument has evolved from a one page manually completed document designed to identify whether homes were fit for human habitation, to a very detailed 20+ page physical property inspection involving a great deal of digital validation. Surveys now include some form of face-to-face questionnaire used to provide a social context to the physical inspection data. Each home nation conducts housing condition surveys and has their own methodological approach to them; some include a social element specifically on housing, some are linked to a more general social survey, some are continuous, some are periodic, some are sporadic! However, the goal is the same; to understand the housing stock and housing circumstances of those living in each nation in order to provide an evidence base for change/improvement. This presentation will: chart how the housing stocks, circumstances and survey methodologies have changed over half a century; highlights the impact of how the data has been used (to improve things); and ask: where next for the UK Housing (Condition) Surveys?

## **10.2 Contributed - Official Statistics & Public Policy: 50th Anniversary of UK National Housing Surveys**

**Thursday 6 September 2pm - 3pm**

### **English Housing Survey at 50: how housing and survey methodologies have changed over half a century**

Brendan Donegan, Reannan Rottier  
*MHCLG*

English Housing Survey at 50: how housing and survey methodologies have changed over half a century  
The English Housing Survey is a national survey of people's housing circumstances and the condition and energy efficiency of homes in England. In 2017, the survey celebrated its 50th birthday. Much has changed since 1967 when the focus of the survey was on whether homes had a bath or shower, an indoor WC, a wash hand basin, and hot and cold water at three points. At that time, 25% of homes lacked one or more of these basic amenities. Some 2.5 million homes didn't have an inside WC. The survey methodology has also changed. For example, the survey instrument has evolved from a one page document designed to identify whether homes were fit for human habitation, to a 23 page physical survey and a 35 minute face-to-face questionnaire. This presentation charts how housing and survey methodologies have changed over half a century, and asks: where next for the English Housing Survey?  
Authors and affiliations: Brendan Donegan[i], Reannan Rottier, Stephanie Freeth, Anna Carlsson-Hyslop and Chauncey Glass, Ministry of Housing, Communities and Local Government (Brendan and one other team member will present).[i]  
Author in the first 10 years of their career

## **10.2 Contributed - Official Statistics & Public Policy: 50th Anniversary of UK National Housing Surveys**

**Thursday 6 September 2pm - 3pm**

### **The House Condition Survey and the impact on policy in Northern Ireland**

Karly Greene

*Northern Ireland Housing Executive*

The Northern Ireland Housing Executive (NIHE) was established in 1971 as the central housing authority for Northern Ireland. The Housing Order in 1981 gave NIHE the statutory responsibility to “regularly examine housing conditions and need” and may “conduct or promote research into any matter relating to any of its functions”. One of the key ways we have performed our statutory duty is through the NI House Condition Survey from 1974. In 1974, one-fifth of dwellings were legally unfit for human habitation and N. Ireland had one of the worst housing stock in Europe. In 2016, 12 surveys later, that figure is now 2%. 40% of dwellings lacked basic amenities, were unfit and in disrepair at that time, now the figure stands 3.5%. Policy and decision makers across N. Ireland have used the NI House Condition Survey to target areas of fuel poverty, improve energy efficiency, boost housing standards through private sector grants and monitor trends in housing tenure and unfitness levels. This vital survey has stood the test of time through our troubled past and now with a different set of challenges in a political stalemate including proposed changes to methodology without a minister, economic challenges and budget cuts. This presentation charts how housing and survey methodologies have changed over almost 50 years and how the information has been used to improve housing stock.

### 10.3 Contributed - Applications of Statistics: Surveys and Censuses

Thursday 6 September 2pm - 3pm

#### **Application of the Hierarchical Hybrid Bayes to a small-area estimation problem using survey and census data from Mexico**

Hector Najera

*University of Bristol*

Context: The hierarchical Bayes (HB) method has very attractive properties for small area estimation (SAE), such as precision, flexibility and computational efficiency. However, the HB might not be a feasible option for high-dimensional problems, i.e. complex models with very large samples. For example, the Mexican SAE project, relied on three methods: the HB model, an ad hoc adaptation of the Elbers, Lanjouw and Lanjouw method and an the Empirical Best Linear Unbiased Predictor (EBLUP). The HB failed to reproduce the direct-design based estimates for the 32 states and provided unstable results for the circa 2450 municipalities. The Hamiltonian (Hybrid) Monte Carlo (HMC) computation is a recent breakthrough in Bayesian estimation that improves further the advantages of HB models and has not been yet applied to a real-data SAE problem. This raises a perfect opportunity to assess whether the Hybrid Hierarchical Bayes (HHB) could improve the estimates from the HB approach.

Objectives: To produce municipal-level estimates of poverty using a large survey, illustrate the advantages of the HMC for complex SAE problems and show how it can be applied to other contexts (Tonga data example).

Method: The presentation draws on a HB estimator (3-level hierarchical model) but relies on the HMC and not on the standard MCMC approach- it uses a HHB estimator. Data: Survey data from Mexico 2010 and 2015 (circa quarter of a million) and a sample of the population Census (circa 10 million cases) from Mexico.

Results: The HBB not only was much quicker than the HB but was capable of reproducing the direct-design based estimates for the 32 states for both years. The HBB also produced very accurate predictions of a variable generated from the Census. The municipal-level estimates were highly consistent with the EBLUP and the ELL method. The HHB is easy to reproduce in other contexts (Data from Tonga).

## 10.3 Contributed - Applications of Statistics: Surveys and Censuses

Thursday 6 September 2pm - 3pm

### Design-based "optimal" calibration weights under unit nonresponse in survey sampling

Per Gösta Andersson  
*Stockholm University*

High nonresponse is a very common problem in sample surveys today. In statistical terms we are worried about increased bias and variance of estimators for population quantities such as totals or means. Different methods have been suggested in order to compensate for this phenomenon. We can roughly divide them into imputation and calibration and it is the latter approach we will focus on here. A wide spectrum of possibilities is included in the class of calibration estimators. We explore linear calibration, where we suggest using a nonresponse version of the design-based optimal regression estimator. Comparisons are made between this estimator and a GREG type estimator. Distance measures play a very important part in the construction of calibration estimators. We show that an estimator of the average response propensity (probability) can be included in the "optimal" distance measure under nonresponse, which will help reducing the bias of the resulting estimator. To illustrate empirically the theoretically derived results for the suggested estimators, a simulation study has been carried out. The population is called KYBOK and consists of clerical municipalities in Sweden, where the variables include financial as well as size measurements. The results are encouraging for the "optimal" estimator in combination with the estimated average response propensity, where the bias was highly reduced for the Poisson sampling cases in the study.

## 10.3 Contributed - Applications of Statistics: Surveys and Censuses

Thursday 6 September 2pm - 3pm

### Automated cleaning of large administrative data-sets

Andrew Sutton, James Bowsher-Murray, Will Perks  
ONS

Following the recommendations of the Bean review, we are now at the advent of a new era in which massive data-sets are being used in the production of key economic statistics. Whilst this brings with it many advantages, especially in terms of data completeness, there are still challenges around how we can detect and potentially edit anomalies in such big data-sets. Clearly traditional manual data cleaning, as has been applied to surveys and small administrative data-sets in the past, is no-longer feasible, and a new, automated solution is now needed. In this talk we describe methods that we have been developing for automated anomaly detection and editing in large administrative data-sets. We demonstrate their practical application using the case-study of HMRC VAT returns data. VAT returns data is currently being used in the production of short-term output indicators, and we have recently been investigating its potential for estimating intermediate consumption for use in supply-use balancing. As such, VAT data is a key bellwether for proving such novel techniques and there are potentially much wider applications.

## 10.4 Contributed - Social Statistics

Thursday 6 September 2pm - 3pm

### Predicting the Results of the 2017 UK General Election

Timothy Martyn Hill

LV

This is the latest in a series considering the effectiveness of metrics at elections. This covers the 2017 UK General Election and is the companion to Significance articles published in 2017 on that election. Metrics are created by others and used to predict the outcome the 2017 UK General Election. So the question arises: which ones are the best? To answer this question we list the common metrics (polls, odds, models), we note the different level of detail for each metric (two candidates, three candidates, four candidates), and the methods we will use to measure their accuracy. We then use the chosen methods to compare those metrics to the final results, both within classes (which poll the best poll, which model the best model...), between classes (which is better: poll, model, odds...), and over time (which is better the day before, the month before, six months before...) Finally we present the conclusions.

See also\* <https://www.significancemagazine.com/politics/555-forecast-error-predictors-of-the-2017-uk-general-election>\* <https://www.significancemagazine.com/files/GE2017-final-with-appendices.pdf>

## **10.4 Contributed - Social Statistics**

**Thursday 6 September 2pm - 3pm**

### **Quantifying the Bradley effect: an application to US and UK elections**

John Fry

*Manchester Metropolitan University*

In this paper we discuss the issue of systematic bias in Opinion Polls caused by Socially Desirable Response - SDR – a phenomenon variously known as the Bradley Effect or the Whitley Effect. In quantifying the analogy between financial and political systems we show that the net result is that polls over-price the probability of Socially Desirable Responses. The effect is known to provide a good explanation of recent referenda results. Applications to the latest US and UK elections are also discussed.

## 10.4 Contributed - Social Statistics

Thursday 6 September 2pm - 3pm

### **Hierarchical model for forecasting the outcomes of binary referenda**

Arkadiusz Wisniowski<sup>1</sup>, Jakub Bijak<sup>2</sup>, Jonathan J. Forster<sup>2</sup>, Peter W. F. Smith<sup>2</sup>

<sup>1</sup>*University of Manchester*, <sup>2</sup>*University of Southampton*

In this talk, we propose a statistical model to forecast outcomes of binary referenda based on opinion poll data acquired over a period of time. We demonstrate how the model provides a consistent probabilistic predictions of the final outcomes over the preceding months, effectively smoothing the volatility exhibited by individual polls. We use a Bayesian hierarchical model to capture the dynamics of the opinion polls. Share of the votes in the polls are assumed to be sampled from a multinomial distribution with overdispersion and a possibility of polling company bias. Next, logit-transformed probabilities of voting for Yes, No and those Undecided are assumed to follow a stationary Ornstein-Uhlenbeck process. We illustrate the method using opinion poll data published before the Scottish independence referendum in 2014, in which Scotland voted to remain a part of the United Kingdom, and subsequently validate it on the data related to the 2016 referendum on the continuing membership of the United Kingdom in the European Union.

## 10.5 Contributed - Methods & Theory: Causal inference

Thursday 6 September 2pm - 3pm

### Combining multiple observational data sources to estimate causal effects

Peng Ding,<sup>1</sup> Shu Yang<sup>2</sup>

<sup>1</sup>*University of California, Berkeley*, <sup>2</sup>*North Carolina State University*

The era of big data has witnessed an increasing availability of multiple data sources for statistical analyses. As an important example in causal inference, we consider estimation of causal effects combining big main data with unmeasured confounders and smaller validation data with supplementary information on these confounders. Under the unconfoundedness assumption with completely observed confounders, the smaller validation data allow for constructing consistent estimators for causal effects, but the big main data can only give error-prone estimators in general. However, by leveraging the information in the big main data in a principled way, we can improve the estimation efficiencies yet preserve the consistencies of the initial estimators based solely on the validation data. The proposed framework applies to asymptotically normal estimators, including the commonly-used regression imputation, weighting, and matching estimators, and does not require a correct specification of the model relating the unmeasured confounders to the observed variables. Coupled with appropriate bootstrap procedures, our method is straightforward to implement using software routines for existing estimators.

## 10.5 Contributed - Methods & Theory: Causal inference

Thursday 6 September 2pm - 3pm

### Selection bias in Instrumental Variable analyses

Rachael Hughes, Neil Davies, George Davey Smith, Kate Tilling  
*University of Bristol*

Background: Study participants are rarely a true random sample of the population they are intended to represent, and both known and unknown factors can influence selection of participants. Failure to account for selection in an instrumental variable (IV) analysis may lead to bias. We review the circumstances in which a two stage least squares (2SLS) IV analysis is biased by selection and illustrate the effects of selection bias.

Methods: We use directed acyclic graphs (DAGs) to depict assumptions about selection and show how DAGs can be used to determine when selection bias occurs. Using simulations, we assess the magnitude of the selection bias of the 2SLS estimate, and coverage of its 95% confidence interval (CI) for a range of selection scenarios. We repeat the simulation study for (1) moderate and strong instruments (partial R<sup>2</sup> 0.04 and 0.39 respectively), (2) linear and non-linear treatment-instrument associations, and (3) causal and non-causal treatments.

Results: The 2SLS estimate is unbiased and CI coverage is close to 95% when selection does not depend on any factors, and when selection only depends on the instrument or only depends on a confounder (of the outcome-treatment association). However, the 2SLS estimate is biased with poor CI coverage if selection: 1) depends on treatment and/or outcome, 2) depends on treatment and instrument, 3) depends on instrument and confounder, or 4) depends on treatment and confounder. Decreasing the instrument strength results in an increase in the level of bias for selection partly depending on the instrument but has little effect on the bias of the other selection scenarios. Changing the treatment-instrument association from linear to nonlinear reduces the size of the standard errors, but its effect on the bias depends on the structure of the selection bias.

Conclusion: Selection bias can have a major effect on an IV analysis. Statistical methods are needed for estimating causal effects from non-random samples.

## 10.5 Contributed - Methods & Theory: Causal inference

Thursday 6 September 2pm - 3pm

### Data-adaptive doubly robust instrumental variable methods for treatment effect heterogeneity

Karla Diazordaz<sup>1</sup>, Rhian Daniel<sup>2</sup>, Noemi Kreif<sup>3</sup>  
<sup>1</sup>LSHTM, <sup>2</sup>Cardiff University, <sup>3</sup>University of York

We consider the estimation of the average treatment effect in the treated as a function of baseline covariates, where there is a valid (conditional) instrument. We describe two doubly robust (DR) estimators: a locally efficient g-estimator, and a targeted minimum loss-based estimator (TMLE). These two DR estimators can be viewed as generalisations of the two-stage least squares (TSLS) method to semi-parametric models that make weaker assumptions. We exploit recent theoretical results that extend to the g-estimator the use of data-adaptive fits for the nuisance parameters. A simulation study is used to compare standard TSLS with the two DR estimators' finite-sample performance, (1) when fitted using parametric nuisance models, and (2) using data-adaptive nuisance fits, obtained from the Super Learner, an ensemble machine learning method. Data-adaptive DR estimators have lower bias and improved coverage, when compared to incorrectly specified parametric DR estimators and TSLS. When the parametric model for the treatment effect curve is correctly specified, the g-estimator outperforms all others, but when this model is misspecified, TMLE performs best, while TSLS can result in huge biases and zero coverage. Finally, we illustrate the methods by reanalysing the COPERS (COping with persistent Pain, Effectiveness Research in Self-management) trial to make inference about the causal effect of treatment actually received, and the extent to which this is modified by depression at baseline.

## **10.6 Contributed - Communicating Statistics: Communicating statistics in a post-truth age**

**Thursday 6 September 2pm - 3pm**

### **Understanding our impact**

Martin Nicholls, Miles Fletcher  
ONS

Communication impact is notoriously hard to measure, with vanity metrics often reigning supreme over performance. The Office for National Statistics (ONS) is committed to eliminating these from its management information and to only rely on insight that is of corporate and strategic importance; for example, if it is helpful and relevant to society by informing news as well as generating it. Since July 2017, ONS's Communication Division has been working closely with Prime Research, one of the world's leading media intelligence agencies, to develop a series of impact measures that assess the daily media performance and to distil the insight that can come from these. This insight is being used across the business to inform communication planning, improve the understanding of statistics in the news and to ensure our key messages have the gravitas to meet audience expectation. This is your opportunity to learn about what a good performance in the media looks like, to understand some of the communication challenges that we face and the opportunities that we all have to improve the impact of our communications.

## 10.6 Contributed - Communicating Statistics: Communicating statistics in a post-truth age

Thursday 6 September 2pm - 3pm

### Fake stats: mythmaking in the digital media age

Fraser Nelson

*The Spectator*

As a journalist, I frequently encounter myths that worry people for no reason. Polls show most young people think they will earn less than their parents: untrue. Thomas Piketty made a book arguing that  $r > g$  and produced figures showing wealth inequality in Britain was worsening. Untrue (as the IMF and FT later demonstrated). But as Mark Twain didn't quite say, fake stats will make their way around the world before the truth gets its boots on. Once, journalists had the time and resources to counter this. Now, we don't. When politicians spin with words, they can be shot down. When they add a decimal point to their spin, it's dutifully recorded. If the ONS has not published research on an opinion on a disputed subject, others can write on this blank page - knowing they will never be countered. The ONS doesn't publish much on the hot topic of generational income, nor does the UK have an asset distribution series predating 2006-08 Wealth & Assets Survey. Britain has, in the ONS, a trusted adjudicator in statistical disputes. But there are big gaps in its coverage, leaving space for myths to take root. And in a social media era, the myths travel terribly fast.

## 10.6 Contributed - Communicating Statistics: Communicating statistics in a post-truth age

Thursday 6 September 2pm - 3pm

### Perceptions vs reality: Presenting statistics to bust myths

Iain Bell

*Office for National Statistics*

The objectives of the talk are to: Outline the challenges of communicating statistics in the era of fake news and “post-truth” Describe how ONS and government statisticians are changing how they present statistics, and engage with statistics users, to encourage more accurate media reporting and better public understanding of statistics Explain how new technologies and digital services can help deliver trustworthy statistics directly to policy makers and citizens The methods covered will be: Providing examples of where public perception does not match the reality across various policy areas, including migrant population, smoking, drinking and drug addiction, and crime rates. Outline the steps the ONS and the Government Statistical Service are taking to improve presentation and communication of statistics, including: reviewing bulletins to provide a more rounded/balanced narrative, data journalism and data visuals, regular blogging to get messages out or shut down inaccurate coverage, public forums, theme days, publishing articles with regional analysis, and seconding staff to Full Fact and the BBC. Designing new tools and products to support different audiences and widen our direct reach and impact. The launch of new Code of Practice for Official Statistics, and an appetite from non-government bodies to adopt its principles to give their statistics more trust/authority. What the future holds and next steps, for example: opportunities to enrich our data through the Digital Economy Act, Customise My Data and Alexa. Working with social media companies committing to tackle fake news on their platforms, and using software to target influencers on social media. The results and conclusions will be: Statisticians in government and beyond need to think differently about how to reach their intended audiences Digital platforms and the increasing importance of technology and smart devices in daily lives provide opportunity to reach out to citizens The challenge is great, but by thinking differently about how reach different audiences, truth can prevail

## 10.7 Contributed - Data Science

Thursday 6 September 2pm - 3pm

### Traffic flow as an early indicator for GDP

Edward Rowland

*Office for National Statistics*

GDP growth is a key metric in considering the economic status of the UK. The ONS produces quarterly growth figures as preliminary estimates that are typically released several weeks after the end of the quarter from data that does not cover the whole quarter. This delay, and later revisions, present issues for policy makers, investors and businesses as the statistics they use to make informed decisions are not current. Therefore, it would be advantageous to produce early indicators of GDP growth that can reliably detect changes in advance of the GDP figure publication. Here we investigate the potential of using traffic flow data to provide early indication of GDP change. We take annual average daily flow figures and show that they lead annual GDP figures by one year using cross correlations, a trend that is consistent across different vehicle types. And that the previous year's traffic flow correlate with annual GDP growth (Pearson's  $R = 0.76$ ,  $p < 0.01$ ,  $n = 12$ ). This shows that traffic flow is a potential early indicator of GDP growth and compares well with similar work carried out by Stats Netherlands using road sensor data (Killan, Ros, "Road Traffic Correlations with Economic Variables: The Big Data Perspective, 2017). Results from ongoing work will also be included. This uses traffic counts, recorded from camera positions on Motorways and Major A-roads from Highways England that provides traffic flow data at fifteen-minute intervals by site . This data contains counts separated by vehicle size so different types of vehicle (Car, HGVs & Coaches etc.) making it an ideal data-set for further investigating the potential of traffic flow as an early indicator for GDP growth as well as what types of vehicle and at what times and locations (e.g. cars at rush hour on weekdays in the South East) might be the best indicator of economic activity.

## 10.7 Contributed - Data Science

Thursday 6 September 2pm - 3pm

### **Pretty vacant: text analysis of 15 million job vacancies shows the UK labour market in unprecedented detail and can help to solve the UK's productivity puzzle**

Arthur Turrell

*Bank of England*

We use a new dataset of around 15 million job adverts originally posted online to take a look at the UK labour market in unprecedented detail, finding significant heterogeneity across regions and occupations. We map these 'big', naturally occurring data on vacancies into official ONS classifications using text analysis, and then match them to existing survey data on the labour force. We use the matched data to examine whether unwinding occupational and regional 'mismatch' between workers and job vacancies would have boosted productivity and output growth in the UK's post-crisis, 'productivity puzzle' period. We find that unwinding regional mismatch would have substantially boosted output and productivity relative to their realised paths. In a second application of our data, we use a number of supervised and unsupervised machine learning techniques applied to the text of the job vacancies and find a data-driven classification for jobs which cuts across wage, sector, region, and occupation. This classification automatically identifies traditional job roles but also surfaces careers not apparent in current taxonomies of jobs. We show that these machine learning derived groups have explanatory power for variables such as offered wages. In a strong test of external validity, we apply the same groupings to survey data on the supply of labour, and find they also have explanatory power for agreed wages. The methodology developed could be deployed to create instant, data-driven taxonomies in conditions of rapidly changing labour markets and demonstrates the potential of unsupervised machine learning in economics and statistics.

## 10.7 Contributed - Data Science

Thursday 6 September 2pm - 3pm

### Using geospatial data to analyse UK port and shipping operations

Ioannis Tsalamanis, Christopher Bonham, Sonia Williams  
ONS

The Automated Identification System (AIS) is used by ships at sea to quickly and accurately track the movement of other vessels. Recent developments have allowed AIS data to be extracted and analysed offline without the need for dedicated AIS equipment. This talk will present the work undertaken by the Data Science Campus (ONS) to import, process and apply AIS data. A novel unsupervised segmentation will be presented that classifies ship behaviour into one of a number of temporal states. This allows shipping activity to be explored at an aggregate level and port activity to be quantified. Finally, a machine learning approach will be discussed that uses the ship state segments and other AIS data to predict ship delays in and around UK ports.

## 10.8 Contributed - Methods & Theory: High-dimensional regression

Thursday 6 September 2pm - 3pm

### Predictor Variable Prioritization in Nonlinear Models: A Genetic Association Case Study

Seth Flaxman<sup>2</sup>, Lorin Crawford<sup>1</sup>, Daniel Runcie<sup>3</sup>, Mike West<sup>4</sup>

<sup>1</sup>Brown University, <sup>2</sup>Imperial College London, <sup>3</sup>UC Davis, <sup>4</sup>Duke

The central aim in this paper is to address variable selection questions in nonlinear and nonparametric regression. Motivated by statistical genetics, where nonlinear interactions are of particular interest, we introduce a novel, interpretable, and computationally efficient way to summarize the relative importance of predictor variables. Methodologically, we develop the "ReIATive cEntrality" (RATE) measure to prioritize candidate genetic variants that are not just marginally important, but whose associations also stem from significant covarying relationships with other variants in the data. We illustrate RATE through Bayesian Gaussian process regression, but the methodological innovations apply to other nonlinear methods. It is known that nonlinear models often exhibit greater predictive accuracy than linear models, particularly for phenotypes generated by complex genetic architectures. With detailed simulations and an *Arabidopsis thaliana* QTL mapping study, we show that applying RATE enables an explanation for this improved performance.

## 10.8 Contributed - Methods & Theory: High-dimensional regression

Thursday 6 September 2pm - 3pm

### Locally adaptive and wavelet regressions for compositional data

Andrej Srakar

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

Regression for compositional data has been so far largely considered only from a parametric point of view. Aitchison (1982) and Hijazi and Jernigan (2009) modelled regression of a compositional response on a real predictor assuming, as distribution for residuals, the Dirichlet or the logistic-normal distributions. For the same problem, Tolosana Delgado and Van Den Boogart (2011) and Egozcue et al. (2012) proposed a linear model using the coordinates of the response, allowing ordinary least squares theory on the space of coordinates. Latterly, some work adapted non-parametric regression to non-Euclidean manifolds. For example, Di Marzio et al. (2013) pursue the circular case, and Di Marzio et al. (2014) the spherical one. The idea is to develop an intrinsic approach to get readily applicable methods without transforming data via link functions. In a recent article, Di Marzio, Panziera and Venieri (2015) extended this to nonparametric situations, introducing local constant and local linear smoothing for regression with compositional data and treating the cases when either the response, the predictor or both of them are compositions. In our analysis, we extend their analysis to locally adaptive estimators, in particular Haar wavelets and adaptive regression splines. We present a detailed statistical and mathematical elaboration and analysis, some comparison (simulation) results with the performance of other existing estimators for regression with compositional data, while, finally, applying the results to two case studies from economics – inference for inequality indices and international trade.

## 10.8 Contributed - Methods & Theory: High-dimensional regression

Thursday 6 September 2pm - 3pm

### Bayesian regularisation from stochastic constraints

Joshua Bon,<sup>1</sup> Berwin Turlach<sup>1</sup>, Kevin Murray<sup>2</sup>, Christopher Drovandi<sup>3</sup>

<sup>1</sup>*School of Mathematics and Statistics, University of Western Australia*, <sup>2</sup>*School of Population and Global Health, University of Western Australia*, <sup>3</sup>*School of Mathematical Sciences, Queensland University of Technology*

Regularisation in Bayesian modelling uses classes of priors which encourage shrinkage on posterior distributions. This property can be beneficial for sparse or underdetermined problems, and reduce overfitting (with benefits for prediction). In this paper, we propose a probabilistic interpretation for regularisation. We augment a given prior distribution with a stochastic constraint that probabilistically restricts the support of the prior, emitting a regularised prior distribution as a result. This introduces the notion of Bayesian regularisation as an operator that acts on a prior, rather than classes of priors which are considered to have desirable shrinkage properties. The framework of stochastic constraints accommodates regularisation of informative priors, multiple simultaneous regularisation, shrinkage towards subsets and subspaces, and shrinkage towards differential equations. It also opens up new computational possibilities. Regularisation from stochastic constraints generalise some prominent priors in the literature including scale mixtures of normal distributions, such as the horseshoe prior [Carvalho et al., 2010] and are applicable to parametric and nonparametric models. We demonstrate the methodology with splines and Gaussian processes, and show that stochastic constraints regularisation can be easily added to existing software (for example the Stan probabilistic programming language).

## 10.9 Contributed - Environmental & Spatial Statistics

Thursday 6 September 2pm - 3pm

### **Towards a general theory for preferential sampling: detecting the preferential selection of sites over time from a fixed population of possible locations**

James Zidek,<sup>1</sup> Gavin Shaddick<sup>2</sup>

<sup>1</sup>*University of British Columbia*, <sup>2</sup>*University of Exeter*

This paper presents a model for the joint distribution of random spatio-temporal process and an associated indicator random variable that indicates the sites at each time point at which monitors have been placed to measure the process. By embedding these processes in a spatio-temporal framework, the model is able to retrospectively assess the impact on inferences about the probability distribution of the spatio-temporal process as well as about predictions made of the process at site locations during offline years. But the embedding also allows for the modeling of the dynamic selection process itself. Thus in a case study involving particulate air pollution over the UK, the paper presents evidence of significant site-selection bias through time, namely that as time progressed the online sites became ever-less representative of the population of all site locations and were typically relocated to locations with the highest pollution levels. Additional site-selection factors, including perhaps those associated with the administrative decisions behind the selection of the sites, are investigated using the model. A major contribution in the paper is the use of the integrated Laplace approximation (INLA) technique on a discrete mesh grid that allows the fitting of the joint model in a reasonable time, without the loss of much accuracy. The implications of preferential sampling for public policy making as well as environmental health risk assessment are explained.

## 10.9 Contributed - Environmental & Spatial Statistics

Thursday 6 September 2pm - 3pm

### Modelling the spatial extent and severity of extreme European windstorms

Paul Sharkey,<sup>1</sup> Jonathan Tawn<sup>2</sup>, Simon Brown<sup>3</sup>

<sup>1</sup>*JBA Consulting*, <sup>2</sup>*Lancaster University*, <sup>3</sup>*Met Office*

Windstorms are a primary natural hazard affecting Europe that are commonly linked to substantial property and infrastructural damage. Extreme winds are typically generated by extratropical cyclone systems originating in the North Atlantic, which are often characterised by a track of local vorticity maxima. While there have been numerous statistical studies on modelling extreme winds, little has been done to model the influence of the extratropical cyclone on the wind speeds that they generate. By modelling the development of windstorms in a Lagrangian frame of reference, we can assess the joint risk of severe events occurring at multiple sites. In this talk, we present a novel approach to modelling windstorms that preserves the physical characteristics linking the windstorm and the cyclone track by exploring the dependence structure of these characteristics in a Lagrangian frame of reference. We explore a combined copula/spatio-temporal filtering approach to identify and extract the spatial footprint of extreme windstorm events, before using a Markov process to propagate the characteristics of the footprint in time relative to the cyclone track. Our model allows simulation of synthetic windstorm events, which one can use to quantify the risk associated with previously unobserved events at different sites, thus representing a useful tool for practitioners with regard to risk assessment. In particular, we show, for case studies in the northwest of England and eastern Germany, that the spatial extent of windstorms become more localised as its magnitude increases, while our model captures the varying degrees of spatial dependence at different sites.

## 10.9 Contributed - Environmental & Spatial Statistics

Thursday 6 September 2pm - 3pm

### Modelling the spatio-temporal dependence of wind fields over complex terrain

Rachael Quill

*School of Mathematical Sciences, University of Adelaide*

The prediction of wind fields can be framed as the multivariate prediction of wind speed and wind direction across a network. Such prediction must take into account the temporal and spatial correlations within, and between, series observed at each location. Throughout the literature, many examples can be found of hourly wind speed analysis relating to wind energy production over spatial scales of multiple kilometres. However, in applications such as bushfire prediction, wind fields must be predicted in short time intervals over 100's of metres. On these scales, auto-correlation and cross-correlation in wind speed can vary significantly from that previously studied and can often be dynamic. In the context of bushfire modelling, wind direction (and its variability) is also just as influential as wind speed, but there are few studies considering the prediction of wind direction across the landscape. This research aims to statistically model both wind speed and wind direction over the small spatio-temporal scales relevant to bushfire modelling. In order to achieve this, the spatio-temporal dependence structure of the network must be captured. Multivariate autoregressive methods combined with correlation and copula techniques are used to model wind characteristics across case studies located in complex terrain. Novel techniques are employed to account for the circular nature of wind direction, and the multi-modal form of its dependence structures. The dynamic nature of wind speed and direction dependencies is addressed through a partition of the data in terms of distinct weather regimes; more advanced techniques are an active area of research. Finally, with an aim to model wind at scales relevant to real-time bushfire prediction, the methods developed within this study must be fast and scalable. A number of parameter reduction techniques, within the construction of multivariate spatio-temporal models, are also considered.

## 11.1 Medical: Simplifying the complex questions in survival analysis

Thursday 6 September 3.30pm - 4.50pm

### Pseudo observations in relative survival estimation

Maja Pohar Perme, Klemen Pavlič

*University of Ljubljana, Faculty of Medicine*

Pseudo-observations present a general approach that can be used for estimation in survival analysis. The idea is to replace the possibly censored survival times by an outcome that is defined for all individuals at all follow-up times despite censoring. In this way, the problem of censoring is removed and one can proceed with standard analyses for non-censored data. In this talk, we focus on the use of pseudo-observations in relative survival estimation. Several estimators have been proposed for net survival estimation, but only recently, a consistent estimator has been introduced. Its use in practice has revealed an excessively large variance when estimating net survival of older age groups. We first simplify the problem by considering a non-censored case to show that the problem of large variance is intrinsic to the definition of net survival and not a property of a specific estimator. We then continue from the definition of net survival and generalize it to the censored case by the use of pseudo-observations. The estimator developed in this way has all the desired properties, we also provide a formula for its variance. Since pseudo-observations are available in several statistical packages, this new estimator is easy to implement. It has several interesting theoretical properties, its main advantage in practice is the fact that it does not require numerical integration. This also implies it can be directly used with life-table data, i.e. data grouped in intervals of time. We illustrate the properties of our proposal with simulations and a real data example of colon cancer patients.

## **11.1 Medical: Simplifying the complex questions in survival analysis**

**Thursday 6 September 3.30pm - 4.50pm**

### **The measurement of socioeconomic inequalities in life years lost by cause of death**

Aurélien Latouche

*Institut Curie*

Quantifying health inequalities in absolute term is of prime interest for decision making and inter-countries comparison. Yet, absolute inequalities using either rates or hazards do not translate into a time dimension, making their interpretation difficult for policy-makers. The Slope Index of Inequality (SII) was recently formalized and we propose an extension of the SII to the expected number of life years lost before an upper age as well as its decomposition by cause of death. The methodology is illustrated in a representative 1% sample of the French population. The SII in life years lost is easily understandable and its decomposition of the all-cause SII attributable to a given cause provides a sound estimation of the burden of a given cause of death on absolute health inequalities.

## 11.1 Medical: Simplifying the complex questions in survival analysis

Thursday 6 September 3.30pm - 4.50pm

### Solving the Fine-Gray riddle

Hein Putter<sup>1</sup>, Jan Beyersmann<sup>2</sup>, Martin Schumacher<sup>3</sup>, Hans C. van Houwelingen<sup>4</sup>

<sup>1</sup>Leiden University Medical Cent, <sup>2</sup>Institute of Statistics, Ulm University, <sup>3</sup>Institute for Medical Biometry and Statistics, Faculty of Medicine and Medical Center—University of, <sup>4</sup>Department of Biomedical Data Sciences, Leiden University Medical Center

The Fine-Gray proportional subdistribution hazards model has been puzzling many people since its introduction. The main reason for the uneasy feeling is that the approach considers individuals still at risk for competing risk 1 after they fell victim to risk 2. The subdistribution hazard and the extended risk sets, where subjects who failed of the competing risk remain in the risk set, are generally perceived as unnatural. One could say it is somewhat of a riddle why the Fine-Gray approach yields valid inference. To take away these uneasy feelings we explore the link between the Fine-Gray and cause-specific approaches in more detail. We introduce the reduction factor as representing the proportion of subjects in the Fine-Gray risk set that has not yet experienced a competing event. In the presence of covariates, the dependence of the reduction factor on a covariate gives information on how the effect of the covariate on the cause-specific hazard and the subdistribution rate relate. We discuss estimation and modeling of the reduction factor, and show how they can be used in various ways to estimate cumulative incidences, given covariates. Methods are illustrated on data of the European Blood and Marrow Society (EBMT).

## 11.2 Official Statistics & Public Policy: A new model for publishing GDP

Thursday 6 September 3.30pm - 4.50pm

### A new model for publishing GDP

James Scruton<sup>1</sup>, Sumit Dey-Chowdhury<sup>1</sup>, Gemma Tetlow<sup>2</sup>, Amit Kara<sup>3</sup>

<sup>1</sup>Office for National Statistics, <sup>2</sup>Institute for Government, <sup>3</sup>NIESR

The session will showcase the work that the UK Office for National Statistics (ONS) has been doing to radically change the way that GDP estimates are published and communicated. From July 2018, ONS introduced a new publication model that presents a more coherent picture of the economy. In summary, this model gives two (rather than three) estimates of quarterly GDP, and speeds up the publication of the Index of Services publication by 2 weeks, enabling the publication of monthly GDP estimates. While our current publication model has the benefit of quick and credible estimates of GDP, the proposed approach results in estimates of monthly GDP as well as allowing ONS to publish a quicker balanced picture of the whole economy and a higher quality first estimate of GDP. The move of the Index of Services also gives further clarity to the publication of economic statistics and was the final key step in reducing the “see-saw” narrative that can emanate from statistics on a related theme being published at different times. It is also noteworthy that the publication of monthly GDP estimates consolidates the UK’s position at the forefront of short term GDP estimation. The conference will be held shortly after the new model is introduced, and we will be able to present on the following areas:

Presentation 1: James Scruton, Head of Monthly GDP, Office for National Statistics

- The rationale and the need for change
- Communication of the changes: developing new products and communicating with different audiences

Presentation 2: Sumit Dey-Chowdhury, Lead GDP economist, Office for National Statistics

- The benefits and trade-offs of the new model
- Interpreting monthly estimates of GDP

Presentation 3: Gemma Tetlow, Institute for Government

- A media response to the new model – how do users view the changes?

Presentation 4: Amit Kara, NIESR

- A policy response to the new model – how do the changes meet policy needs?

### **11.3 Applications of Statistics: Biofortified food crops in developing countries: are they effective?**

**Thursday 6 September 3.30pm - 4.50pm**

#### **Statistical issues related to dietary intake as the response variable in intervention trials**

Ruth Keogh

*London School of Hygiene and Tropical Medicine*

The focus of this talk is dietary intervention trials. We will explore the statistical issues involved when the response variable, intake of a food or nutrient, is based on self-report data that are subject to inherent measurement error. There has been little work on handling error in this context. A particular feature of self-reported dietary intake data is that the error may be differential by intervention group. Measurement error methods require information on the nature of the errors in the self-report data, and we assume that there is a calibration sub-study in which unbiased biomarker data are available. I will outline methods for handling measurement error in this setting and use theory and simulations to show how self-report and biomarker data may be combined to estimate the intervention effect. Methods are illustrated using data from the Trial of Nonpharmacologic Intervention in the Elderly, in which the intervention was a sodium-lowering diet and the response was sodium intake.

### **11.3 Applications of Statistics: Biofortified food crops in developing countries: are they effective?**

**Thursday 6 September 3.30pm - 4.50pm**

#### **Statistical challenges in assessing the effectiveness of biofortification**

Ian Plewis

*University of Manchester*

Biofortification is, in essence, an agricultural intervention that depends for its success on farmers' willingness to grow the biofortified crops and on consumers' willingness to pay for them. The principle of its effectiveness can be established through RCTs at a local level and, more generally, from meta-analyses of these trials. But if biofortification is to be seen to make a difference at a population level, for example to those indicators of child health (e.g. the prevalence of stunting) proposed to determine whether the UN Sustainable Development Goals are being met, then more data and different methods of analysis are needed. The paper will focus on these issues and will draw on the way evidence has been constructed for another agricultural intervention – the introduction of genetically engineered cotton in India – to suggest ways forward.

### **11.3 Applications of Statistics: Biofortified food crops in developing countries: are they effective?**

**Thursday 6 September 3.30pm - 4.50pm**

#### **Food Fortification and Biofortification in Low and Middle-Income Countries: Global Status and Evidence of Impact**

Greg Garrett

*Global Alliance for Improved Nutrition (GAIN)*

This presentation will summarize the global status of implementation of food fortification and biofortification. This will include an overview of where these two population-based interventions have been delivered; what scale, coverage and impact has been achieved through these strategies to date; and a summary of the primary research and programmatic gaps moving forward. The presentation will also aim to unpack some of the emerging complementarities between large-scale food fortification and biofortification.

## **11.4 Social Statistics: Surveys, Bayes and the Machine Learning Craze - the future of forecasting elections**

**Thursday 6 September 3.30pm - 4.50pm**

### **In defence of surveys; why polling is a sampling problem**

Gary Brown

*Office for National Statistics,*

Expert sampling knowledge is used to review current polling methods with critique and recommendations for improvement. Issues such as survey design, response and coverage bias are discussed, and techniques are suggested for improving the representativeness of poll samples and their reliability.

## **11.5 Methods & Theory: Regression & dimension reduction methods for high dimensional data**

**Thursday 6 September 3.30pm - 4.50pm**

### **Near-equivalence of Dimension Reduction Methods in Large Panels of Macro-variables**

Efstathia Bura,<sup>1</sup> Alessandro Barbarino<sup>2</sup>

<sup>1</sup>*Vienna University of Technology*, <sup>2</sup>*Federal Reserve Board*

In an extensive pseudo out-of-sample horserace, classical estimators (dynamic factor models, RIDGE and partial least squares regression) and the novel to forecasting Sliced Inverse Regression and Recurrent Neural Networks exhibit almost near-equivalent forecasting accuracy in a large panel of macroeconomic variables across targets, horizons and subsamples. This finding motivates our theoretical contributions in this paper. We show that most widely used linear dimension reduction methods solve closely related maximization problems with solutions that can be decomposed in signal and scaling components. We organize them under a common scheme that sheds light on their commonalities and differences as well as on their functionality.

## 11.5 Methods & Theory: Regression & dimension reduction methods for high dimensional data

Thursday 6 September 3.30pm - 4.50pm

### Dimension reduction for functional data based on weak conditional moments Functional dimension reduction

Bing Li, <sup>1</sup>Jun Song<sup>2</sup>

<sup>1</sup>*The Pennsylvania State University*, <sup>2</sup>*University of North Carolina at Charlotte*

We develop a general theory and estimation methods for functional linear sufficient dimension reduction, where both the predictor and the response can be random functions, or even vectors of functions. Unlike the existing dimension reduction methods, our approach does not rely on the estimation of conditional mean and conditional variance. Instead, it is based on a new statistical construction --- the weak conditional expectation, which is based on Carleman operators and their inducing functions. Weak conditional expectation is a generalization of conditional expectation. Its key advantage is to replace the projection on to an L<sup>2</sup>-space -- which defines conditional expectation -- by projection on to an arbitrary Hilbert space, while still maintaining the unbiasedness of the related dimension reduction methods. This flexibility is particularly important for functional data, because attempting to estimate a full-fledged conditional mean or conditional variance by slicing or smoothing over the space of vector-valued functions may be inefficient due to the curse of dimensionality. We evaluated the performances of the our new methods by simulation and in several applied settings.

## 11.7 Data Science: Evolution of Data-Centric Engineering

Thursday 6 September 3.30pm - 4.50pm

### Machine learning of beam structures by blending physical models and data

Alastair Gregory<sup>1</sup>, Din-Houn Lau<sup>1</sup>, Liam Butler<sup>2</sup>

<sup>1</sup>Imperial College London, <sup>2</sup>University of Cambridge

In many engineering applications such as structural health monitoring, structures are instrumented with sensors that produce large quantities of data. This data is frequently combined with machine learning techniques to infer information about the structures. Physical and finite element models of these structures are also used to model their responses; often parameters within these models are updated using the data. This blending of data and physics is an example of 'data-centric engineering'. We consider the case of instrumented sleepers on a railway bridge. This talk presents the application of a novel machine learning technique based on Gaussian processes to implement this type of data and physics blending. The technique uses a Gaussian process prior informed by physics, for the inference of sleeper curvature where data can be provided from a fibre-optic sensor network. The benefits of combining the data and physics in this way include the estimation of material properties and a tool for damage detection.

## 11.7 Data Science: Evolution of Data-Centric Engineering

Thursday 6 September 3.30pm - 4.50pm

### **A Statistical Model for Addressing Uncertainty in the Assessment of Performance of Ageing Retaining Structures to Groundwater Inundation**

Victoria Stephenson,<sup>1</sup> Chris Oates<sup>2</sup>

<sup>1</sup>*UCL/The Alan Turing Institute*, <sup>2</sup>*Newcastle University/The Alan Turing Institute*

Increased rainfall amounts in the UK are leading to a growing recurrence of groundwater flooding events. These cause problems for embedded and retaining structures through heightened and fluctuating pore water pressures, which lead to ground deformation and subsequent movement and damage to construction. These issues are especially significant in relation to ageing structures, where decay in the engineered system and subsequent loss of structural integrity exacerbates damage caused by groundwater effects. Uncertainty associated with the various components of the problem further compounds the risk. As a result, understanding and managing the risks for engineering structures in this context is complex, often requiring estimation of geotechnical parameters that lead to assumptive measures of performance. Presented here is an analysis of groundwater conditions surrounding a tunnel retaining structure in central London. The groundwater model is then used for the assessment of an historic, ageing retaining wall to failure by sliding and overturning. A statistical approach is also taken to the definition of the key engineering parameters in the mechanical models defining the sliding and overturning failure mechanisms, including those of the wall and surrounding ground. The final output of the work is a statistically derived model for failure of an idealised retaining structure in the context of uncertain geotechnical and groundwater conditions. The work outputs a solution which can be both robustly computed and referenced against a specific geographical location. The objective of the work is to describe and evidence a simple technique for overcoming uncertainty in groundwater flood risks to retaining structures, and through this promote the use of larger datasets for the improvement of parameter quantification in the future.

## **11.9 Environmental & Spatial Statistics: Geospatial methods for global health applications**

**Thursday 6 September 3.30pm - 4.50pm**

### **Mapping changes in housing in Africa, 2000 to 2015**

Samir Bhatt

*Imperial College London*

Adequate housing is a fundamental human right and an important objective of Sustainable Development Goal 11, but housing conditions across Africa have not been comprehensively described. We mapped changes in housing in Africa from 2000 to 2015 by combining national survey data within a Bayesian statistical framework. Our analysis revealed a dramatic transformation of housing in sub-Saharan Africa between 2000 and 2015, with the prevalence of improved and modern housing doubling from 11% to 23%. However, lived in slum housing in 2015. Our maps provide a novel and powerful mechanism to track housing trends in Africa and guide interventions in the post-2015 development era.

## **11.9 Environmental & Spatial Statistics: Geospatial methods for global health applications**

**Thursday 6 September 3.30pm - 4.50pm**

### **Mapping multiple diseases and risk factors: practical lessons from diagnosing modeled spatiotemporal health predictions**

Aaron Osgood-Zimmerman

*Univ. of Washington, Institute for Health Metrics and Evaluation*

For over a decade, the Global Burden of Disease (GBD) study at the Institute for Health Metrics and Evaluation (IHME) has been working to produce national estimates of burden for over three hundred different diseases and additional health risk factors at the global scale. Three years ago, the Local Burden of Disease (LBD) team was started to resolve national level estimates for the five leading causes of children-under-five mortality to the much finer 5x5 km resolution. Since then, we have added other diseases, risk factors, and useful covariates to our mapping efforts. During this time, one of our greatest challenges has been to properly vet, diagnose, iterate, and improve our spatio-temporal health maps. In practice, we have seen that many of the classic metrics for goodness of fit and model comparison may not differentiate well between predictive maps that may be quite meaningfully different to the global health scientists and policy makers who would use them. In this talk, I'll describe some of the problems and limitations with model diagnostics and comparisons that we have encountered, as well as the visual and numerical tools that we have implemented and developed to improve and further our mapping efforts.

## **11.9 Environmental & Spatial Statistics: Geospatial methods for global health applications**

**Thursday 6 September 3.30pm - 4.50pm**

### **Age-structured modelling of population and childhood vaccination coverage in low and middle income countries**

Victor Alegana

*University of Southampton*

The long-term goal of the global effort to tackle infectious diseases is national and regional elimination and eventually eradication. This requires disaggregated estimates or fine scale multi-temporal mapping to help guide planning and resource distribution. This presentation will outline some of methods being used for modelling interventions in childhood diseases. Using selected case studies, the presentation will demonstrate efforts on mapping infectious vaccination coverage to combat measles in children. Classical approaches include incorporating effects of environment as well as measuring healthcare accessibility. The fine-scale maps depict areas that should be prioritised by governments in low-and middle-income countries to target resources to specific local populations. This is particularly important for diseases being targeted for elimination.

## **Keynote 7 – Champion (President’s Invited) Lecture**

**Thursday 6 September 3.30pm - 4.50pm**

### **Viewed through Tinted Glasses? Public Reactions to the Brexit Process**

Sir John Curtice

*University of Strathclyde*

'When the facts change, I change my mind.' That is supposedly what a wise economist should do. Doubtless many statisticians would agree. Trouble is, we may not agree on the facts, a point illustrated by the row during the EU referendum about the claim that leaving would free up £350m a week that could be spent on the NHS. Since then the process of negotiating Britain's withdrawal from the EU has thrown up lots more 'facts' - but perhaps they have simply been viewed through a partisan lens that means their import is disputed.

How have voters reacted to the 'facts' that have emerged in the two years since the EU referendum? Have they shown a willingness to change their minds? Or has new 'information' simply been interpreted so that it conforms to voters' prior predispositions - with the result that Britain remains as divided as ever about Brexit?

## ***POSTER PRESENTATIONS***

## **Evaluating temporal patterns of snakebite in Sri Lanka: the potential for higher snakebite burdens with climate change**

Dileepa Ediriweera

*University of Kelaniya, Sri Lanka and Lancaster University, UK*

**Background:** Snakebite is a neglected tropical disease that has been overlooked by healthcare decision makers in many countries. Previous studies have reported seasonal variation in hospital admission rates due to snakebites in endemic countries including Sri Lanka, but seasonal patterns have not been investigated in detail.

**Methods:** A national community-based survey was conducted during the period of August 2012 to June 2013. The survey used a multistage cluster design, sampled 165 665 individuals living in 44,136 households and recorded all recalled snakebite events that had occurred during the preceding year. Log-linear models were fitted to describe the expected number of snakebites occurring in each month, taking into account seasonal trends and weather conditions, and addressing the effects of variation in survey effort during the study and of recall bias amongst survey respondents.

**Results:** Snakebite events showed a clear seasonal variation. Typically, snakebite incidence is highest during November to December followed by March to May and August, but this can vary between years due to variations in relative humidity, which is also a risk-factor. Low relative humidity levels are associated with high snakebite incidence. If current climate change projections are correct, this could lead to an increase in the annual snakebite burden of 31.3% (95% CI: 10.7 – 55.7) during the next 25 to 50 years.

**Conclusions:** Snakebite in Sri Lanka shows seasonal variation. Additionally, more snakebites can be expected during periods of lower than expected humidity. Global climate change is likely to increase the incidence of snakebite in Sri Lanka.

## **Symbolic input-output analysis: a harmonic analysis approach to combining statistical distributions**

Andrej Srakar

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

We provide a new, stochastic, approach to study input-output analysis and calculation of multipliers. We apply the findings to the calculation of production and employment multipliers for selected European countries. Input-output (IO) analysis is, in principle, one of the most commonly used, but a non-stochastic approach to national accounts. Yet, it suffers from several common critiques, not least being fixed input structure in each industry; all products of an industry are identical or are made in fixed proportions to each other; and each industry exhibits constant returns to scale in production. To this end, we use symbolic data analysis (following e.g. Michalski, Diday and Stepp, 1981; Diday, 1987; Cazes et al., 1997; Brito, 1994; 1995; Billard and Diday, 2000; 2002; 2003; 2006; Ichino, 2011; Verde and Irpino, 2015; Diday, 2015) to construct distributions (histogram variables) in the cells of IO tables instead of numerical aggregated values. Using such approach, we are able to include the stochastic component in the modelling with IO tables in a novel way. To combine the cells/distributions we provide foundations of a symbolic Leontief distribution calculus, based on harmonic analysis (in particular, the concept of convolutions, see Hardy, 1949; 1966; Grattan-Guinness, 1970; Bottazzini, 1986; Luzin, 1998; Heil, 2010). We derive the confidence intervals of production and employment multipliers, calculated in a novel way, and apply the methodology to the EU countries in the period 2008-2011. Preliminary results confirm the validity of the approach and show important advantages of taking into account the stochastic component of the IO analysis in a manner as proposed in the paper. In conclusion, possibilities of solving the usual limitations of IO analysis using the new approach are addressed, although future work is needed to explore this promising path of future work in this field.

## **A Modified Generalized Chain Ratio in Regression Estimator**

Olaniyi Mathew Olayiwola,<sup>1</sup> Adebisi Apantaku<sup>2</sup>, Adebisi Olayiwola<sup>3</sup>, Jaiyeola Opeyemi<sup>2</sup>

<sup>1</sup>*Federal University of Agriculture, Abeokuta, Nigeria*, <sup>2</sup>*Funaab, Nigeria*, <sup>3</sup>*Oyo State Teaching Commission, Funaab, Nigeria*

Generalized Chain ratio in regression type estimator is efficient for estimating the population mean. Many authors have derived a Generalized Chain ratio in regression type estimator. However, the computation of its Mean Square Error (MSE) is cumbersome based on the fact that several iterations have to be done, hence the need for a modified generalized chain ratio in regression estimator with lower MSE. This study proposed a modified generalized chain ratio in regression estimator which is less cumbersome in its computation. Two data sets were used in this study. The mean square errors in the existing and proposed estimators were derived and relative efficiency was determined. The results of this study showed that the proposed estimator gave lower MSE for the two data sets, hence it is more efficient.

## **Developing a Goodness of Fit Test for a High Dimensional Multilevel Binary Model**

Marina Roshini Sooriyarachchi, Gayara Fernando  
*University of Colombo, Sri Lanka*

Before making inferences about a population from a fitted model, it is necessary to determine whether the fitted model describes the data well. A model that well describes the data at hand is said to be adequately fitting the considered data. A goodness of fit test determines the model adequacy of a fitted model. Some recent studies have shown the necessity of goodness of fit tests for high dimensional binary multilevel models in determining the model adequacy due to unavailability of satisfactory goodness of fit tests that specifically look at this aspect. Traditional GOF tests for single level will not work here as now the observations are correlated. The main objective of this research was to develop a goodness of fit test to determine the model adequacy of high dimensional binary multilevel models within the framework of the specialized software MLwiN. The theories behind single level Hosmer and Lemeshow test, Lipsitz et al.'s GOF test for ordinal data, method by Perera et al. for binary two level multilevel models and limited information GOF testing concepts by Maydeu-Olivares et al. were considered as the bases for coming up with the novel goodness of fit test. To determine whether type I error and power hold for the newly developed GOF test, extensive simulations were carried out considering a three level random intercept only model by varying the cluster and ICC combinations at the existing levels of the hierarchy. The results obtained for the simulations suggested that type I error holds for all the considered cluster and ICC combinations except for the cluster combinations having small sizes at all levels. A high power was obtained for the test under the alternative model considered for all the considered cluster and ICC combinations.

## **TWO-STAGE SAMPLING INVOLVING PROBABILITY PROPORTIONAL TO SIZE (PPS) METHOD**

Adetola Konku,<sup>1</sup> O. A. Olorojo<sup>1</sup>, O. r. Olaoye<sup>1</sup>, Adebisi Olayiwola<sup>2</sup>

<sup>1</sup>*Federal University of Agriculture, Abeokuta*, <sup>2</sup>*Oyo Tescom*

This study focused on application of Probability Proportional to Size (PPS) sampling method in two-stage sampling. Secondary data were used. The first preliminary sample of size ten (10) clusters was selected independently by Simple Random Sampling Without Replacement (SRSWOR). The Lahiri's method was used to select First Stage Units (FSU) of Six (6) clusters from the preliminary sample using Probability Proportional to Size with Replacement (PPSWR). Within each selected First stage clusters, sub-samples of Second Stage Units (SSU) were selected with SRSWOR. An estimator under PPS was derived and the expressions for the Bias and Mean Square Error (MSE) were obtained. The results showed that the PPS enhances the efficiency of the derived estimator. The derived estimator is therefore preferred in the estimation of a heterogeneous population parameter.

## **STATISTICAL ANALYSIS OF DETERMINANTS OF NIGERIA GDP**

Kazeem Adekunle Oyekunle<sup>1</sup>, Oladipupo Adeleke<sup>2</sup>, Abosede Adeniran<sup>3</sup>, Rebecca Samuel<sup>4</sup>  
<sup>1</sup>*Department of Statistics, Federal University of Agriculture Abeokuta, Nigeria,* <sup>2</sup>*Oyo East Local Government,* <sup>3</sup>*Ekiti State University, Ado-Ekiti,* <sup>4</sup>*Crawford University, Ogun State, Nigeria*

Nigeria is classified as a mixed economy emerging market, and has already reached middle income status according to the World Bank, with its abundant supply of natural resource. Authors were of the views that most developing countries, agriculture has been assigned an important role in national development. This research work examined the contributions of Agriculture, Oil(Export and Re.Export), External-Reserves, N Exchange-rate, Transportation, Education and Communication to Nigeria economy. Principal component analysis was used to examine the contributions of the variables. The result showed that as Agriculture, Oil (Export and Re.Export), Transportation, Education and Communication increases there was an increase in GDP while there was a decline in external reserve and exchange rate.

## **New Confidence Interval Estimator of the Signal-to-Noise Ratio Based on Asymptotic Sampling Distribution**

Ahmed Najeeb Albatineh<sup>1</sup>, Ibrahimou Boubakari<sup>2</sup>, B M golam Kibria<sup>2</sup>

<sup>1</sup>*Kuwait University*, <sup>2</sup>*Florida International University*

In this paper, the asymptotic distribution of the Signal-to-Noise Ratio (SNR) is derived and a new confidence interval for the SNR is introduced. An evaluation of the performance of the new interval compared to Sharma and Krishna (S-K) (1994) confidence interval for the SNR using Monte Carlo simulations is conducted. Data were randomly generated from normal, log-normal, chi square, Gamma, and Weibull distributions. Simulations revealed that the performance of S-K interval is totally dependent on the amount of noise introduced and that it has a constant width for a given sample size. The S-K interval performs poorly in four of the distributions unless the SNR is around one. It is recommended against using the S-K interval for data from log-normal distribution even with SNR=1. Unlike the S-K interval which does not account for skewness and kurtosis of the distribution, the new confidence interval for the SNR outperforms S-K for all five distributions discussed, especially when SNR $\geq$ 2. The use of ranked set sampling (RSS) instead of simple random sampling (SRS) improved the performance of both intervals as measured by coverage probability.

## **Time series modelling of wind power using the inflated beta distribution**

Fraser Tough

*Renewable Energy Systems*

Synthetic wind power time series are frequently utilised within renewables, feeding into other energy system models such as sizing battery storage systems or when considering grid curtailment. A typical approach when generating synthetic power is to split power into discrete categories, assuming a multinomial distribution for power category. First, second or third order Markov matrices are then estimated after rendering the series stationary by removing trend. Power can then be simulated over a discrete state space via Monte Carlo approaches, retaining the diurnal, seasonal and autoregressive nature of power. The reason that power is discretised is that it is bounded by 0 and the rated power of the turbine – simulating across categories guarantees that the output will also be bounded. However, power is not categorical, it is continuous. A common workaround is to model wind speed using standard time series approaches, transforming simulated wind speed to power via the theoretical power curve. However, the theoretical power curve rarely describes observed power due to other factors such as temperature. Here, a new time series approach is presented, allowing modelling of power directly over a continuous rather than discrete state space. Power varies over time, with highest densities at 0 and the rated power due to the nature of the power curve. Once scaled, the distribution is well described by the inflated beta distribution. However, the parameters of the distribution vary by hour and month. An inflated beta regression model is proposed where the model parameters can be modelled as functions of lagged power, month, hour and any other variables that the power distribution may depend on. This novel approach was applied to a real-world problem - to generate a power time series for a wind farm within Scotland. Coupled with availability simulations based on the same framework, energy based availability losses by time of day and time and year could be quantified.

## **Bayesian forecasting of subnational population change**

Arkadiusz Wisniowski<sup>1</sup>, James Raymer<sup>2</sup>

<sup>1</sup>*University of Manchester*, <sup>2</sup>*Australian National University*

In this talk, we extend the well-known multiregional population projection model developed by Andrei Rogers and colleagues to be fully probabilistic. The projections are based on forecasts of age- and sex-specific fertility, mortality, interregional migration, immigration and emigration for eight states and territories of Australia. We extend and apply bilinear models, such as the well-known Lee-Carter model used for forecasting mortality, as well as the log-linear models used for capturing patterns cross-tabulations of demographic variables classified by various dimensions such as age, sex, country of origin, education. The innovation of this article is combining of the two approaches to deal with the high dimensionality of the demographic components. We demonstrate how the method permits taking into account the correlation structure across age, sex and regions in the demographic forecasting and thus provide a robust modelling platform for projecting subnational populations with measures of uncertainty.

## Graphical Principles Cheat Sheet

Andrew Wright, Mark Baillie, Baldur Magnusson, Andrew Wright, Ruquan You, Julie Jones, Marc Vandemeulebroecke  
*Novartis*

**Objectives:** The goal of this poster is to summarize Good Graphical Principles for statistical graphics. The poster is accompanied by a single-page “Cheat Sheet” handout, and it is related to a separate presentation that illustrates these principles in a live demonstration.

**Methods:** A large body of literature and training material including [1] - [7] has been condensed into a single-page reference sheet. The various principles have been grouped into sections such as: selecting the right base graph; an effectiveness ranking of graphical attributes (volume, color hue, depth, area, angle, length etc.); facilitating comparisons; the use of color; enhancing legibility and clarity; various implementation considerations; and a checklist for the most important of these aspects. Each point is illustrated concisely and intuitively with thumbnail graphs.

**Results:** We present and share a carefully designed “Graphical Principles Cheat Sheet” for every-day use in graphical data exploration and the production of graphics for communicating analysis results and conclusions. The sheet is available as a hand-out with the poster.

**Conclusions:** A carefully designed single-page reference sheet on Good Graphical Principles is a useful tool for the creation of clear and impactful graphics. We want to share this work for the benefit of a wider audience.

*References:*[1] Cleveland (1985). *The elements of graphing data*. New York: Chapman and Hall.[2] Duke, Bancken, Crowe, Soukup, Botsis, Forshee (2015). *Seeing is believing: Good graphic design principles for medical research*. *Statistics in Medicine* 34 (22), 3040-3059.[3] Krause, OConnell (editors, 2012): *A picture is worth a thousand tables*. New York: Springer. [4] Robbins (2013). *Creating more effective graphs*. Chart House. [5] Tufte (2001). *The visual display of quantitative information*, 2nd ed. Cheshire, CT: Graphics Press.[6] Tukey (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley. [7] Wong (2013). *The Wall Street Journal Guide to Information Graphics: The Dos and Don'ts of Presenting Data, Facts, and Figures*. WW Norton.

## **Global methods to reconcile bilateral trade data asymmetries**

Marilyn Thomas, Katie O'Farrell

ONS

The poster presentation covers the main causes of bilateral trade data asymmetries and the methods adopted globally to reconcile them. There are three types of approach to tackling asymmetries: (1) The use of common reporting frameworks which are aimed at standardising the measurement of trade. These frameworks have recently been harmonised, especially for trade in services, but many lower-income countries have limited resources to implement new frameworks. (2) Bottom-up reconciliation exercises between trading partners. These are resource-intensive, bilateral exercises undertaken by pairs of countries. (3) Top down approaches to reconciliation (the 'Global Model' approach) include the use of mirror data from partner countries, or estimates of bias in asymmetries giving rule-of-thumb indications and possible adjustments. These include the adjustment applied by the OECD and Eurostat, and a new top-down approach investigated by ONS and Thomas Baranga (2018).

## Stability Analysis of Stochastic Model for Stock Market Prices

Iyai Davies, Amadi Uchenna, Roseline Ndu

*Department of Mathematics, Rivers State University, Port Harcourt, Rivers State , Nigeria*

In this paper, new differential equation model that could impact the expected returns of investors in stock exchange market has been considered with a stochastic volatility in the equation, the unstable nature of the stock price changes in the stock market are analyzed using data from Nigeria stock exchange. The results were obtained by developing stochastic vector differential equation, exploring the properties of the fundamental matrix solution (a function of the drift) in the equation and by placing continuity condition on the stochastic part (a function of the volatility). Analysis on the effect coefficient of correlation  $\xi$  and the volatility  $\sigma$  on the stability of system showed that;  $\xi$  denotes the sources of randomness for the underlying Weiner process and the volatility. It captures the leverage effect affecting the size of the tails; the skewness of the return distribution. That is, if  $\xi < 0$ , then volatility increases and asset price return decreases, thus, leading to the spread of left tail and squeeze in right tail of the distribution, thus, creating a fat left-tailed distribution. If  $\xi > 0$ , volatility increases with increase in asset price return. This causes the right tail to spread with a squeeze in the left tail of the distribution, thus, creating a fat-tailed distribution. If  $\xi = 0$ , the skewness is close to zero. The effect of  $\sigma$  is mainly on the peak of the distribution. When  $\sigma = 0$ , the volatility becomes deterministic because the processes will be zero and  $\xi$  also zero. This will lead to a normal distribution of stock returns as in the Black Scholes model. Theoretically, increase in  $\sigma$  also increases the peak, creating fatty tails on both sides implying that increase in market volatility  $\sigma$  gave higher peak of the distribution and vice versa

## **Identifying clusters of bacteria from SNP information**

Alasdair Noble

*AgResearch NZ*

26 Soils were collected from around New Zealand and 10 isolates of *Rhizobia leguminosarum* bacteria were collected from each soil. These isolates were sent for DNA sequencing to find SNP's that explained the variation between the isolates. It is expected that many of the isolates would be from a few common strains as soils are generally inoculated with rhizobia when clovers are sown. However it is hoped that there may be some agriculturally interesting strains amongst them. A small sample of the strains along with replicated copies of three commercial strains were resequenced to estimate the random variability in the sequencing so we can identify groups of the same strain and some "new" individuals. In this talk I will explain why identifying these new strains could be important for agriculture and how I separated the strains into the groups.

## **Mining signals of astronomical sources via Bayesian nonparametric mixture modelling**

Andrea Sottosanti, Mauro Bernardi, Alessandra Rosalba Brazzale  
*University of Padua*

The search of gamma-ray sources in the extra-galactic space is one of the main targets of the Fermi LAT collaboration, which aims to identify and study the nature of high energy phenomena in the Universe. This requires to separate their signal from the diffuse gamma-ray background over the entire area observed by the telescope. From a statistical perspective, we can account for these two phenomena using a mixture of two densities, where the first models the spread of photons around the source, while the second represents the background contamination. In this work, we propose a novel approach to the identification of gamma-ray sources using a Dirichlet Process mixture, which we combine with a flexible Bayesian nonparametric model based on B-splines to account for the irregular shape of the background. The reference model is hence a mixture of two Dirichlet Process mixtures, which will allow us to discover and locate a possible infinite number of clusters in the map. From the results obtained on a region observed by the Fermi LAT we can conclude that the proposed approach achieves both, the estimation of the number of sources present in the map and a complete separation of their signal from the background.

## **Statistical considerations when designing an efficacy study for a consumer ovulation test**

Lorrae Marriott, Graham Warren, Sarah Johnson  
*SPD Development Company Ltd*

With the large variety of products to help women achieve pregnancy available it is important that the performance measures provided are of the highest quality to enable consumers to make the best decision when choosing a test. An efficacy study provides additional confidence to a consumer on the benefit of buying a particular product. There are very few published studies in this area therefore little guidance on what a “good” study looks like. With our many years of experience in consumer testing we have brought together our best practices to design an efficacy study which will produce high quality unbiased results. This paper will discuss, with a review of previous studies, the important factors to be considered when setting up an efficacy study: Defining the outcomes - identification of pregnancy  
Length of study – How many menstrual cycles is a subject recruited for  
How to identify and recruit suitable subjects – inclusion/exclusion criteria  
Manage subject drop out and pre-trial pregnancies  
How to reduce potential unequal loss to follow up in the cohorts  
Estimation of sample size required based on expected pregnancy rates in the population – interim review of sample size (is the pregnancy rate assumption holding) and how this is managed  
Randomisation and blinding of study  
Guidance on the clear reporting of study results

## **Asymptotic distribution of the bootstrap parameter estimator of an AR(2) model**

Bambang Suprihatin, Endro Setyo Cahyono, Alfensi Faruk

*Mathematics Department, University of Sriwijaya, Indralaya, Indonesia*

This paper is the extension of our research about asymptotic distribution of the bootstrap parameter estimator for the AR(1) model. We investigate the asymptotic distribution of the bootstrap parameter estimator of a second order autoregressive AR(2) model by applying the delta method. The asymptotic distribution is the crucial property in inference of statistics. We conclude that the bootstrap parameter estimator of the AR(2) model converges asymptotically in distribution to the bivariate normal distribution

## **Inferences in Step Stress Partially Accelerated Life Tests for Compound Rayleigh Distribution with Progressive First-Failure Censoring**

Tahani Abushal

*Umm Al-Qura University*

This study considers the problem of estimating the unknown parameters of the compound Rayleigh distribution with progressive first-failure censoring scheme during step-stress partially accelerated life tests (ALT). Progressive first-failure censoring and accelerated life testing are performed to decrease the duration of testing and to lower test expenses. The maximum likelihood estimators (MLEs) and Bayes estimates (BEs) for the distribution parameters and acceleration factor are obtained. The optimal time for stress change is determined. Furthermore, the approximate, bootstrap and credible confidence intervals (CIs) of the parameters are derived. Methods of Markov chain Monte Carlo (MCMC) are used to obtain the Bayes estimates. Finally, the accuracy of the MLEs and BEs for the model parameters is investigated through simulation studies. The simulation performed to compare the suggested methods for several accelerated factors, several parameter values, several sample sizes ( $n, m$ ) and three several CSs. Approximate, credible and bootstrap CIs have been determined for  $\alpha$ ,  $\beta$ , and  $\lambda$ . From the results we realized that: 1. We observed that in most of the cases the length of approximate, bootstrap and credible CIs are decreasing when the sample size is increasing, excepting a little cases; this probably due to the inconstancy in the data. 2. Overall, the MCMC credible CIs of  $\alpha$ ,  $\beta$  and  $\lambda$  yield a good results than approximate and bootstrap CIs for the length of CIs. 3. It may be noted that for fixed observed failures and sample sizes, the first scheme-I, yield lower lengths for the three methods of the CIs in contrast to the another two schemes. 4. It can also be seen that the Bayes estimates of  $\alpha$ ,  $\beta$ ,  $\lambda$  give a good results for the MSEs and RABs than for MLEs in most of the cases considered. In general, as sample size  $m/n$  increases, the MSEs and RABs of MLEs and Bayes estimates of  $\alpha$ ,  $\beta$ , and  $\lambda$  decrease.

## **Multivariate Analysis Advancements and Applications by Subspace-based Techniques**

Xu Huang

*De Montfort University*

Alongside the increasing speed of human society developments and technological advancements, the complexity level of multivariate analysis has rapidly risen due to the prevalence of knowledge and science, regardless of the existing controversial drawbacks of the wide range of empirical methods (parametric and limited nonparametric approaches). This research aims to expand the multivariate extension of subspace-based techniques on multivariate analysis and brings novel contributions to not only the theoretical advancements but also broadening the horizon of the corresponding applications in complex systems like economics and social sciences. Subspace-based techniques adopted in this research include Singular Value Decomposition (SVD), Singular Spectrum Analysis (SSA) and Convergent Cross Mapping (CCM), which all have the advantages of being nonparametric approaches, assumption-free, no limitations to nonlinearity or complex dynamics, signal and noise together as a whole as the research object. This research proposed two novel multivariate analysis methods based on the study of subspace-based techniques: the mutual association measure based on the eigenvalue-based criterion; and the hybrid causality detection approach by combining SSA and Convergent Cross Mapping (CCM). Both simulations and several successful implementations are conducted for the critical evaluation of the proposed advancements with promising robust performances. The proposed approaches offer the interested parties a different angle to resolve the multivariate analysis questions in a reduced form, data-oriented aspect. It is also expected to open the research opportunities of nonparametric multivariate analysis through the advanced, inclusive subspace-based techniques that show strong adaptability and capability in the complex system analysis in economics and social science.

## **Graphical evaluation of the prediction capabilities of composite mixed-resolution designs in spherical regions**

Polycarp Chigbu, Eugene Ukaegbu, Cynthia Umegwuagu  
*University of Nigeria, Nsukka, Nigeria*

**Objective:** In this study, we propose three-dimensional variance dispersion graphs as graphical tools for evaluating the prediction variance performances of some composite mixed-resolution designs in spherical regions. The aim of the study is to provide a graphical method that is capable of providing very useful information about the designs' prediction variance properties throughout the entire design region which the single-value optimality criteria would not easily provide.

**Method:** The graphical procedure was obtained through analytically tedious but tractable matrix algebra involving the prediction variance function. The composite mixed-resolution designs were assessed in the spherical regions by considering the impact of the practical and spherical axial distances on the designs' prediction variance characteristics while replicating the cube and star portions of the designs.

**Results/Conclusion:** The variance dispersion graphs show that augmenting the composite mixed-resolution designs with one and/or two additional star portions, in most cases, improves the designs' prediction capabilities in the spherical region.

## **Newspaper analytics and the dynamics of the exchange rate in Peru**

Luciana Figueroa,<sup>1</sup> Erick Lahura<sup>2</sup>

<sup>1</sup>*Pontifical Catholic University of Peru*, <sup>2</sup>*Central Reserve Bank of Peru*

This poster explores the effects of macroeconomics news on the level and volatility of daily exchange rate in Peru. The data are daily and covers the period 01/01/2014 - 31/03/2018. Macroeconomic news are summarized in several "news indexes" constructed from newspaper articles published online during the times when the central bank does not intervene in the foreign exchange (forex) market. One group of indexes measure the occurrence and intensity of key macroeconomic words related to the forex market. A second group is built by combining individual words into "good news" and "bad news" indexes, using text mining techniques. We use standard linear and nonlinear time-series methods in order to estimate the contribution of macroeconomic news to forecasting exchange rate. The results show that these models can improve their forecasting accuracy by including macroeconomic news.

## **Life and Death of Pixels - Spatial Distribution and Quality Assessment of Dysfunctional Pixels in Digital Detectors**

Julia Brettschneider<sup>1</sup>, Clair Barnes<sup>2</sup>, Jay Warnett<sup>1</sup>, Greg Gibbons<sup>1</sup>, Mark Williams<sup>1</sup>, Tom Nichols<sup>3</sup>, Wilfrid Kendall<sup>1</sup>

<sup>1</sup>*University of Warwick*, <sup>2</sup>*UCL*, <sup>3</sup>*Big Data Institute*

A collection of binary values indexed by a grid can form a model to describe the locations of dysfunctional pixels in a digital detector. Natural questions also arise around the spatial distribution of the dysfunctional pixels and how observed patterns of dysfunctional pixels may be interpreted. After modelling occurrences of dysfunctional pixels as a planar point process we develop a higher level approach for analysing their spatial distributions. Key idea is to move from the notion of a dysfunctional pixel to the concept of a damage event defined by configurations of dysfunctional pixels using a typology based on local grid geometry. High density regions can be detected using density estimation of the damage event process, so remaining areas becomes suitable candidates for complete spatial randomness. This approach decouples observed damage from the detector resolution prescribed by  $\lambda$  and from the exact shape of dysfunctional pixel configurations. We propose a detector quality toolkit that allows users to monitor their technology following these principles. The methods allow users of detector based imaging technologies to detect, distinguish and monitor different types of quality damage and to identify the ones linked to specific causes. We apply our methods to a collection of bad pixel maps obtained as part of regular monitoring routines of a detector used in X-ray computed tomography.

## **An Analysis of Dynamic PET Scans with a Number of Separate Radio tracer Injections**

Fengyun Gu, Francisco Hernandez, Liam O'Suilleabhain, Ran Ren, Jian Huang, Eric Wolsztynski, Finbarr O'Sullivan  
*University College Cork*

An Analysis of Dynamic PET Scans with a Number of Separate Radio-tracer Injections  
.Fengyun Gu, Francisco Hernandez, Liam O'Suilleabhain,Ran Ren, Jian Huang, Eric Wolsztynski and Finbarr O'Sullivan  
Department of Statistics, University College Cork, Ireland  
fengyungu@126.com  
Support: Science Foundation Ireland Grant No. PI-11/1027  
Objectives: P-glycoprotein (P-gp) has a role in the removal of foreign substances out of cells. A series of studies were carried out to evaluate the ability to use positron emission tomography (PET) scanning with a C-11 labeled Verapamil radiotracer [Vp], to image the P-gp status. The studies involved dynamic Vp PET imaging, before and after the administration of cyclosporine, an agent which has a known dramatic effect on the P-gp status. Previous reports on these studies focused on consideration of time-course data from a limited number of regions of interest. Our work reports on a statistical analysis that is capable of producing a comprehensive voxel-level analysis of the full data set.  
Methods: The analysis approach is based on a mixture model that can express the full voxel-level time-course (over all available dynamic scans) as a positive linear combination of underlying (basis) time-courses that are represented in the data. This scheme can be recognized as a form of traditional factor or principal component analysis. An adaptive weighted least squares approach is used for implementation. Kinetic analysis of the basis time-courses enables a voxel-by-voxel mapping of parameters describing the P-gp activity. The analysis is implemented in R.  
Results: Illustrations from humans and in non-human primates studies are presented. Generated 3-D metabolic images demonstrate the ability to map the P-gp activity and its response to the cyclosporine injection. We also present numerical studies demonstrating the reliability/consistency of the methodology.  
Conclusion: Comprehensive analysis of Vp-PET scanning data to recover local P-gp information can be carried out with reliance on statistical mixture models. The approach is found to be promising.

## **Analysing Language Trends Across Time Using Large Text Corpora**

Rachel Carrington

*University of Nottingham*

Text mining is an ever growing area of research interest, due to the large amount of text data now available. Of particular interest in linguistic applications is being able to determine how words change over time, both in meaning and in usage. However, although a variety of methods for modelling text have been developed, much less attention has been given to the time-dependent case. The methods for this that do exist tend to rely on dividing data into broad periods, the choice of which is rather arbitrary, making the results hard to interpret, and risking losing some information. We propose a new method which aims to represent words as vectors which are continuous functions of time, allowing us to make use of all of the time information available. The intention is that words which are more similar to each other will be closer together. By measuring how the distances between words changes across time, we can see how the relationships between these words change, which allows us to infer changes in relative meaning and usage.

## **A Practical Approach for Simulating Noise Characteristics of an Operational PET Scanner**

Jian Huang, Liam O'Suilleabháin, Ran Ren, Tian Mou, Finbarr O'Sullivan  
*University College Cork, Ireland*

Positron Emission Tomography (PET) is widely used in the clinical management of many cancers - in staging, therapy planning and evaluation of therapy response. However, because of low tracer dosages and other reasons, PET images display increased noise levels compared to other modalities such as Computed Tomography (CT) and Magnetic Resonance Imaging (MRI). Improved understanding of the characteristics of PET imaging measurements could lead to improvements in noise suppression. In our previous study (1), using data generated from PET studies on a physical phantom with uniform source distribution and a numerical uniform phantom, we demonstrated that the distributional characteristics of iteratively reconstructed PET images can be well-described by a Gamma model form capturing the non-Gaussian skewness. In this study we consider validating this result on PET studies on non-uniform phantoms. Furthermore, we investigate relationship between the scale parameter of the Gamma model (computing as variance/mean) and the underline source distribution, attenuation pattern, the number of iterations and post-smoothness. Comprehensive simulation studies using range of attenuation maps including ones from patient data were conducted. Simulation results show: Iteratively reconstructed PET images can be well-described by a Gamma model. Higher attenuation will lead to higher the scale parameter values. There is dependence of the scale parameter on the source distribution, but strength of the dependence is impacted by the number of iterations and post-smoothness. The established empirical relationship could lead to a simple and fast approach to simulating bias and variance characteristics of an operational PET scanner, which can be easily implemented in R. SUPPORT: SFI PI 11/1027; NIH R33CA225310.

## **A new X -bar chart for detecting small and large shift in process mean**

Ekele Alih

*Dept. of Maths/Statistics, Federal Polytechnic Idah, Kogi State, Nigeria*

This article proposes a new X-bar chart for controlling the process mean based on repetitive sampling scheme. This chart makes use of two pairs of upper and lower control limits to effectively detect the process mean shift away from the target value for small, medium and large shift. The control limits coefficients,  $k_1$  and  $k_2$  were obtained through Monte-Carlo simulation experiment by minimizing the out-of-control average run length and at the same time maximizing the in-control average run length. The new chart returns to the conventional X-bar chart when  $k_1 = k_2$ . The proposed control chart has a comparative advantage over the synthetic control chart, the exponentially moving average (EWMA) chart, and the joint joint X-bar-EWMA chart in that it can detect both small and large shift in mean while the others work well only when the shift in mean is small. A real-life example is presented.

## Mixture Modelling for AIF Extraction from Dynamic PET Studies

Zhaoyan Xiu<sup>1</sup>, Jian Huang<sup>1</sup>, Fengyun Gu<sup>1</sup>, Janet O'Sullivan<sup>1</sup>, Eric Wolsztynski<sup>1</sup>, David Mankoff<sup>2</sup>, Finbarr O'Sullivan<sup>1</sup>

<sup>1</sup>University College Cork, <sup>2</sup>University of Pennsylvania

Mixture Modelling for AIF Extraction from Dynamic PET Studies  
Objective: Kinetic modelling of dynamic PET data requires knowledge of tracer concentration in blood plasma so-called arterial input function. Arterial blood sampling is the gold standard of the methods used to measure the AIF. However, this method is not favoured in routine clinical practice, as it is highly invasive and labour intensive. Hence, a number of methods have been proposed to accurately extract the AIF directly from image data. We proposed a mixture modelling method for AIF extraction. Methods: The tracer atoms at a sampled blood-site each have a specific history in the body. We represent the history for a specific atom by a travel time. This travel time is modelled as a sum of time ( $T_{ir}$ ) for the atom to initially progress from the injection site to the right ventricle of the heart and time ( $TC$ ) it spends in circulation before being sampled.  $T_{ir}$  is modelled as a realization from a Gamma distribution.  $TC$  is modelled by a mixture model. The Gamma parameters and mixing components of the model are fixed and do not change across subjects, and they are estimated from a collection of arterial sampled data for the tracer on different subjects. The mixing weights are estimated for each subject. A penalized nonlinear least squares process, incorporating a generalized cross-validation score, is proposed for estimate mixing weights. Results: The proposed method is first validated on arterial sampled FDG and H<sub>2</sub>O data. Three components model can fit FDG data well, while two components model can fit H<sub>2</sub>O data well. Both models achieved smaller cross-validation error in comparison to the existing triexponential model and whole body circulation model. Finally, the method is illustrated on image data for H<sub>2</sub>O and FDG. SUPPORT: SFI PI 11/1027; NIH R33CA225310.

## **The History of a Population Pyramid: A Snapshot of the Romanian Demographics**

Aura Popa

*The Statistical Company*

The population pyramid – also called an age structure diagram or an age-sex pyramid – is a graphical illustration and represents the breakdown of the population by gender and age at a given point in time. It consists of two histograms, one for each gender (by convention, men on the left and women on the right; the oldest age group is on top, the youngest one at the bottom) where the numbers are shown horizontally and the ages vertically. The numbers by gender and by age depend on interactions between fertility, mortality and migrations. This poster is an empirical research based on census data displayed through data visualisation of the population pyramid, showing the turbulent history of Romanian demographic evolution. The analysis of the age-sex pyramid will show the impact on the population of the World War I, the World War II, the decree 770 of the communist leader Nicolae Ceausescu in 1966 that abolished the abortion and contraceptives, the fall of the communist regime up to nowadays – an E.U. population with its free movement that resulted in the brain drain effect Romania is experiencing in the last two decades placing it as the second highest emigration growth rate after Syria according to a 2015 UN report.

## **A logistic regression approach for combining likelihood ratio results from different software for DNA mixture interpretation**

Tereza Neocleous,<sup>1</sup> Eugenio Alladio<sup>2</sup>, Paolo Garofano<sup>3</sup>, Marco Vincenti<sup>2</sup>, Dimitra Eleftheriou<sup>1</sup>  
<sup>1</sup>*University of Glasgow*, <sup>2</sup>*Department of Chemistry, University of Turin and Laboratorio di Biologia Forense, Centro Regionale Antidoping e di Tossicologia "A. Bertinaria" of Orbassano (Torino)*, <sup>3</sup>*Laboratorio di Biologia Forense, Centro Regionale Antidoping e di Tossicologia "A. Bertinaria" of Orbassano (Torino)*

DNA typing interpretation plays a critical role in courtrooms. Due to different interpretation procedures, Low-Template DNA (LT- DNA) mixture profiles obtained from the crime scene can be challenging, with their interpretation proving difficult, especially in the presence of random sample degradation and varying quantities of DNA available. Another source of complexity in the interpretation of DNA mixtures is the fact that there are three main models that can be used when interpreting data obtained from electropherograms (epg): the binary approach (currently obsolete), the semi-continuous approach and the fully-continuous approach. These models present different degrees of difficulty in terms of application and interpretation. Despite the forensic community having proposed several recommendations over the past few years, a standardised, "universal" and rigorous approach to LT-DNA mixture analyses has still to be defined. The main aim of this study is to build a generalized, comprehensive approach in order to combine the likelihood ratio (LR) results that are provided by the different probabilistic approaches and corresponding software for DNA mixture interpretation. Several ad hoc DNA 2- and 3-person mixtures, already employed for validation purposes, were analysed at the Laboratory of Forensic Genetics of the Regional Antidoping and Toxicology Center "A. Bertinaria" (Orbassano, Italy) by means of different probabilistic software including both semi-continuous (i.e. Lab Retriever and LRmix Studio) and fully-continuous (i.e. DNA•VIEW® and EuroForMix). Logistic regression approaches were used to combine the different LR values, together with the degradation index and DNA quantification parameters. In this pilot study, logistic regression produces promising results indicating that it could act as a valuable tool for the combination of LRs provided by various biostatistical software and approaches. Furthermore, it allows combination of the various LR results together with parameters such as the degradation index and the overall amount of DNA to be amplified.

## **A MODIFIED ADAPTIVE CLUSTER SAMPLING DESIGN FOR ESTIMATING RARE AND HIDDEN POPULATION SIZE**

Yasir Ademola Osoyode<sup>1</sup>, Saheed Agboluaje<sup>2</sup>, Kester Asugha<sup>1</sup>, Festus Isakunle<sup>1</sup>, Isiaka Tijani<sup>1</sup>, Abdulmateen Olanipekun<sup>1</sup>, Solomon Akinyemi<sup>1</sup>

<sup>1</sup>*Department of Statistics, Federal University of Agriculture Abeokuta, Nigeria,* <sup>2</sup>*The Polytechnic, Ibadan*

Adaptive Cluster sampling is an efficient method for estimating rare and hidden clustered population size. Various authors had worked on estimators in adaptive cluster sampling, but did not consider presence of auxiliary information for the purpose of increasing the precision of the estimator. This study derived Adaptive cluster sampling estimator for estimating the size of rare and hidden population in the presence of auxiliary information. The biasness and Mean Square Error (MSE) for the existing and proposed estimator were determined and compared. Asymptotic confidence interval (CI) for the population mean and total were constructed. The MSE for the derived estimator was lesser compared with the existing one, hence more efficient.

## **Effective methods for replacement of zeroes in lipidomics data**

Simone Cuff, Sven Mickelmann, Valerie O'Donnell  
*Cardiff University*

Lipidomics is a difficult but growing field within immunology and metabolomics which uses high resolution mass spectrometry to separate lipid subtypes from within blood plasma. The thousands of lipid subtypes present can present a lipid fingerprint of patients or cells to determine biological importance of lipids, or potentially diagnostic signatures. The system is extremely sensitive. This is an advantage in that it can pick up many fine differences in biomarkers. However, it is also disadvantage in that small changes in protocols or running conditions can cause apparent changes in the lipids detected. In particular, variations between runs can mean that small environmental changes can lead to the same lipid being mis-identified as two separate lipids in two runs, giving apparent zeroes. This project examines methods of processing and replacing of zeroes to maximise biologically relevant results from large lipidomic datasets. Data consisted of 25,891 features measured across 44 total samples from 3 groups, plus 5 control samples. It found that standard methods of removing samples from the analysis on the basis of the proportion of zeroes resulted in the loss of strongly differential counts and after investigation of a number of alternatives, a two-tier method in which the proportion of zeroes present in each subgroup was also taken into account was found to be a more accurate way to judge cutoff for inclusion or exclusion from the dataset. Using the lipidomics dataset from patients with differing genotypes, we show that this allowed the discovery of biomarkers which were strongly differentially expressed between genotypes but which were otherwise discarded from the analysis.

## **Development of new computational methods to elucidate the molecular dynamics of drug treatment using CyTOF**

Marie Trussart, Terry Speed, Charis Teh, Daniel Gray  
*Walter and Eliza Hall Institute*

A new technology which couples mass spectrometry with metal-conjugated antibodies permit the profiling of cellular phenotype and signalling cell state of millions of individual cells. Specifically, such technological advance with cytometry time of flight CyTOF has successfully enabled a comprehensive panel of surface and intracellular protein markers to unravel complex signalling networks and to delineate cell subsets in heterogeneous tissues such as blood, bone marrow and tumours. Indeed, mass cytometers are able to analyse simultaneously more than 40 unique parameters per sample at the single cell level. While there are already many tools available to analyse the resulting data and answer questions of interest, the development of biostatistical and bioinformatics methods to address all these different topics is still in its infancy. We have already worked on the application of an algorithm enabling correction of signal fluctuations and we are proposing a new computational method that is able not only to handle and remove technical unwanted variation but also to refine the previously developed debarcoding method in high-dimensional mass spectrometry data. As an application of CyTOF to elucidate the effect of drug treatment, our collaborators have been developing a custom-made protocol to better understand the impact of different treatments on Multiple Myeloma (MM) cell lines and Chronic Lymphocytic Leukaemia (CLL) patients. Current approaches to understand cancer heterogeneity do not resolve the protein changes dynamics which occur at the level of individuals cells. By using this new technology CyTOF, we will analyse CLL patient samples and aim to resolve disease-relevant pathways at the single-cell level to better understand the impact of each drug that underlie sensitivity or resistance to the treatment and how its efficacy is influenced by heterogeneity.

## **Hierarchical cluster modeling of asthma and its clinical phenotypes**

Ronald Wesonga, Charles Bakheit, Faisal Ababneh

*Department of Mathematics and Statistics, Sultan Qaboos University, Oman*

Asthma and chronic obstructive pulmonary diseases are assiduous inflammatory diseases with substantial effects on health and well-being of individuals and communities and are known to pose great financial implications. This study was aimed at developing potential clusters of these diseases, based on the reported incidences by the Ministry of Health, Oman. The study mapped the clusters onto the potential asthma clinical phenotypes. Overall, five hierarchical agglomerative clusters (HAC) were developed using the disease incidences for six years (2010-2015) based on the HAC complete linkage. Results indicated that the majority (92%) of asthma patients had relatively easy to control symptoms of asthma. Patients with easy to control asthma symptoms were mainly young and male, while more older female patients experienced difficult to control asthma symptoms. The study recommends complete linkage under the hierarchical agglomerative clustering because it provided better performance than the single and average linkages. Asthma disease clustering facilitates targeted prioritization and management of the disease. Further research is sought to develop cost of illness optimization models for healthcare resource allocation to combat the asthma scourge.

## **Model predictors selection to predict the recovery of arm function post-stroke after three months**

Ahmad Al-shallawi, Dimitra Blana, Anand Pandyan  
*Keele University*

**Introduction:** Prediction recovery of loss arm function post-stroke is complicated, especially in patients with severe levels of initial impairments. Therefore, the existing models are not clinically useful, improving variable selection to reform modelling process can improve model performance. The aim is to: use the K-means clustering method to group and, then, apply the stepwise regression method and adaptive Lasso for logistic regression predictors selection and compare the results.

**Methods:** We used data-set of 176 patients, with baseline measurements (giving 136 independent variables) taken within a week of a stroke, and arm function measurements (Action Research Arm Test – ARAT) taken at baseline, 4 and 12 weeks post stroke (the dependent variable)<sup>1</sup>. K-means was applied to group the three-time measures (3 column) of ARAT's outcome to exclude the patients with high level of function. Data was split into 75% training and 25% testing. Stepwise regression and Adaptive LASSO were used for variable selection and modelling based on ridge regression weighted<sup>2</sup>. Results testing of each method were compared using accuracy, sensitivity and specificity tools.

**Results:** K-means clustering grouped the ARAT into four groups: three included the moderate to severe patients, and the last one was excluded with low severity of (66) patients. Stepwise regression selected (16) and Adaptive LASSO shrunk 136 predictors to five with accuracy (72% and 76%), sensitivity (64% and 82%), specificity (64%, 71%) respectively.

**Discussion:** Selecting relevant predictors can improve the ability and the accuracy of the developed model. The study's results demonstrate that applying the penalised method of selecting the most important predictors for functional recovery arm post-stroke, would improve the model's predictive precision. Consequently, for clinicians, this can be applied in predicting recovery of patients, and assist in managing workload. This method can select a subset of predictors that have a very significant impact in predicting recovery which agrees significantly with those selected clinically

## Seeded MDS with a large number of cases

Xiangyu Meng, Jian Huang, Finbarr O'Sullivan, Liam O'Suilleabháin, Ran Ren, Zhaoyan Xiu  
*University College Cork*

Seeded MDS with a large number of cases  
Abstract: Multidimensional Scaling (MDS) is a generic name for a family of dimension reduction algorithms used to configure points in low dimensional space given a matrix of pairwise distance between them. Although MDS is powerful for visualization 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 with  $N$ . We examine the possibility of a seeded approach to MDS analysis when  $N$  is too large. Our approach consists of two steps: An MDS analysis based on seed sample of  $n < N$  cases is carried out to map seed sample cases in a  $K$ -dimensional space. A seed sample can be selected as cluster centres obtained by clustering analysis of the data. Using the seed sample 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 dissimilarity matrix distances from these cases. The accuracy of the proposed algorithm is evaluated by the mean squared mismatch between the pairwise distance of cases in the mapped space and the dissimilarity distance values, the seed data being used as a reference. The algorithm is implemented in R and demonstrated on real data sets and simulation studies. The memory usage and computation time and accuracy as a function of the seed sample size  $n$  for the given number of cases  $N$  will be reported. The relative performance of different approaches to selection of seed sample will also be reported.

## Survival analysis of genomics profile for cancer patients

Khaled Alqahtani

*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  $\sigma$  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.

## **Africa's Participation in International Statistical Literacy**

Elieza Paul

*International Statistical Literacy Project-Tanzania*

This paper aimed at investigating the real challenges that hinders this very important initiative in their respective countries

Methodology: 53 ISLP-African country coordinators were contacted through emails and the following question was emailed to them "specifically in your country experience or your point of view, What challenges you always face when organizing or thinking of conducting ISLP-Statistical poster competitions? ' what should be done?" 5 out of 53 contacted country coordinators through their email addresses Findings shows that, 18 African countries have either shown interest or participated in International statistical Literacy project since it started, 9 out of 18 countries have only shown interest but haven't participated neither at a national level nor international level.

Discussion: The international statistical literacy project in Africa remains limited and sometimes is taken for granted, in other words, ISLP in Africa is facing some financial difficulties as the findings have outlined in this paper, this is accelerated by low understanding and appreciation of all matters related to statistics and its application among the stakeholders (sponsors) in different African Countries. Lack of volunteerism spirit, you can see Africa is leading when it comes to the number of ISLP- Country coordinators but poorly participating in the competitions. It can be noted that Out of 53 contacted country coordinators few responded and large part of them didn't respond either by less considering the issue or joined with a lot of expectations and have missed them. Lack of political commitment and less efforts shown by National statistical office as hubs for promoting and supporting statistical literacy in African countries.

## **Modelling and Analysis of Simple Pendulum Computer Experiments Using a Support Vector Regression Model**

Kazeem Osuolale<sup>1</sup>, Waheed Yahya<sup>2</sup>, Babatunde Adeleke<sup>2</sup>

<sup>1</sup>*University of Ibadan*, <sup>2</sup>*University of Ilorin*

Computer experiments are techniques commonly adopted in engineering, scientific and technological applications in the modern time. Its flexibility and relevance has given it a wide acceptability than the classical physical experiments when developing computer experiments and building metamodels of computer simulators. Statistical Design of Experiments (DOE) is fast growing since it is applicable to both physical processes and computer simulation models. In this study, an Orthogonal Array-based Latin Hypercube Design, that is, OA (n, m ) LHD was used to develop simple pendulum computer experiments. The simple pendulum computer experiments were conducted based on the OA (49, 3) LHD using a simple pendulum model. The model was used to mimic a simple pendulum experiment that is conventionally performed in the laboratory. A Support Vector Regression (SVR) model was employed as a computer based metamodel to emulate the simple pendulum computer model in order to minimize the required computational efforts in using a computer model and predict the stoppage time of pendulum at untried inputs. The SVR algorithm was adopted for modelling and analysing simple pendulum computer experiments. MATLAB 2015 package was used to implement the SVR algorithm with Gaussian Radial Basis Function (GRBF) used as a kernel function to ensure that the SVR efficiently captures non-linearities in the simple pendulum computer model. The e-insensitive loss function was used with  $e=0.05$  and  $C=8$  to control the support vector regression model. Results from the SVR model indicated that GRBF trained with 77.6% of the experimental runs with zero bias. The fitted SVR model gave predicted values that were close to the simulated test data as shown in Figure 2.

## Performance of improved estimators in stratified settings in the presence of nuisance parameters

Asma Saleh

*UCL*

There is a persisting interest in methods that can deliver optimal estimation and inference from models whose parameter space  $\theta$  can be written in the form  $\theta = (\psi, \lambda)$ , where  $\psi$  is a scalar parameter of interest and  $\lambda$  is a set of nuisance parameters. Particular interest is on stratified settings when both the number of strata-specific nuisance parameters and the stratum sample size increase to infinity. The main challenge with such models is that even basic requirements, like consistency, from vanilla estimation methods, such as maximum likelihood, are not necessarily satisfied. We investigate the performance of the indirect inference in the basic but challenging setting of binary matched pairs model. We derive the asymptotic expansion of the bias of the indirect inference estimator and show that the latter removes the first order bias term of the maximum likelihood estimator. To compare the bias of the two estimators, we implement the simulation-based approach of indirect inference by Kuk (1995) and present results from a small scale simulation study.

## **WHO 0-3 Developmental Indicators - A systematic analysis of child development items from seven assessment tools in ten countries**

Gareth McCray,<sup>1</sup> Gillian Lancaster<sup>1</sup>, Melissa Gladstone<sup>2</sup>, Patricia Kariger<sup>3</sup>, Vanessa Cavallera<sup>4</sup>, Tarun Dua<sup>4</sup>, Magdalena Janus<sup>5</sup>

<sup>1</sup>*Keele University*, <sup>2</sup>*Liverpool University*, <sup>3</sup>*University of California*, <sup>4</sup>*World Health Organization*, <sup>5</sup>*MacMaster University*

Over 200 million children under 5 are not reaching their developmental potential. Currently, no tools exist that can measure children's development in meaningful, valid and culturally/linguistically comparable way. This study aimed to identify items within already validated tools, utilised in low and middle income (LAMI) settings, that have similarly and adequately functioning developmental trajectories to construct as culturally/linguistically unbiased a tool as possible. We located 14 cross-sectional datasets from 10 countries across Africa, Asia and Latin America that used one or more of 7 developmental tools commonly used in LAMI settings. 21,083 children aged 0-3 years from nationally representative samples or those enrolled in large randomized trials were included. However, there was no explicit crossover between the datasets thus, methods combining expert and statistical data had to be devised for linkage. A matrix mapping exercise with Subject Matter Experts (SMEs) identified similar items across tools and item performance statistics were calculated. A consensus process was used to select items for the final tool according to 1) capacity to discriminate by age, and to do so similarly across tools and countries, and 2) their representativeness of developmental indicators identified by previous systematic review. For validation, a 2-Parameter Logistic (2PL) Model with an exponential decay (increasing form) function to model the increase in development with age was fitted using Bayesian methods (Stan). The datasets were linked by 1) 12 common "anchor" item sets, running through the different tools, and 2) the function describing mean change in development with age. A total of 789 developmental items were analysed. A 120-item prototype was created through consensus covering items which showed good cross-cultural/linguistic similarity was drafted. The 2PL model largely reinforced the results of the judgement consensus item selection process. This tool went to be successfully piloted and refined in three countries.

## Non-proportional hazards in randomized clinical trials

John Gregson, Linda Sharples, Jonathan Bartlett, Stuart Pocock  
*LSHTM, Astra Zeneca*

**Background:** Trials of time-to-event outcomes are often planned and analysed using proportional hazards (PH) models. This is not a good choice if the treatment effect deviates from PH. **Objectives:** To classify departures from PH and compare the performance of analysis methods applied to 4 randomised controlled trials (RCTs)

**Methods:** We compared the performance of the following methods: Cox-PH models; restricted mean survival time (RMST); milestone analyses i.e. comparison of proportion with the event at a fixed time-point; accelerated failure time models; and several “combination tests” which test the equality of survival distributions.

**Data:** Four large RCTs in cardiovascular disease each with a distinct pattern of treatment effect: ASCOT-LLA (proportional hazards), ASCOT-BPLA (delayed treatment effect), CHARM (early treatment effect) and EXCEL (diminishing treatment effect, an early treatment effect that attenuates later in follow up).

**Results:** In ASCOT-LLA, where the PH assumption was satisfied, all methods gave significant results, but PH and accelerated failure time models gave smaller p-values. In ASCOT-BPLA, which showed a delayed treatment effect, only PH models, milestone analyses and combination tests yielded significant results (i.e. p-values <0.05). In contrast, in CHARM, which showed an early treatment effect, only RMST and accelerated failure time models and the combination tests yielded significant results. In EXCEL, there was reduced short-term risk with non-surgical treatment compared to CABG surgery, but greater long-term risk. Hazard ratios by year of follow up indicated reduced hazard in the first year, similar hazard in the second year and greater hazard in the third year. Despite neither treatment being clearly superior, two of the three combination tests gave highly significant p-values.

**Conclusions:** If the pattern of non-PH is anticipated, a suitable analysis strategy can be chosen to maximise power. Combination tests detect differences in survival distributions across a range of treatment effect patterns, but their significance does not necessarily indicate superiority of one treatment compared to another.

## **Technologies for a driverless future: How well can cars detect and read road signs?**

Ciaran Ellis, Saket Mohan, Rahul Khatry, Jolyon Carroll  
*Transport Research Laboratory*

To enable driverless cars to use our roads safely, one important task will be the reading of speed limit signs. TRL as part of the MoveUK consortium\*, were asked to evaluate the ability of a speed sign detection system to detect and record the correct speed when passing a sign. Data from real journeys were collected from 4 specially adapted cars. Data from over 200 sensor readings were processed in real-time and transmitted to the Cloud at the end of each journey. It was then up to a statistics and data science team to determine: 1) Whether the cars were detecting road-signs in the correct locations 2) Whether the correct speeds were being read from a mixed-effect binomial logistic regression model was fitted for each question. 6 variables had a significant influence on the likelihood of a detection event in the correct location (including road type and speed of travel). Likewise, 6 variables had a significant influence on the likelihood of a car identifying the speed correctly (the most important being road type and lateral acceleration). In both cases, there was also a large amount of variation explained by the random effect of road sign location. While overall the average probability of detection was high (88%), for some signs the probability of a detection event was very low (minimum of 16%); these differences between signs warrants further investigation. Possible causes of low detection could be unusual road configurations, damaged or obscured signs. Overall the system performed well in a variety of situations, though further work is required to ensure that road sign detection and interpretation are accurate for all normal driving scenarios – this may involve both changes to the algorithm and alternative placement of signs for some road layouts. \*MoveUK project partners include Bosch, Jaguar Land Rover, TRL, Direct Line Group, The Flook and the Royal Borough of Greenwich.

## **A history of political opinion polling in the United Kingdom**

Timothy Martyn Hill

*LV*

The use of polling to predict elections is widespread and excites much public comment at election time. But the public understanding of political polling is limited, with the decades-out-of-date "man with clipboard" still being the public image. We discuss the development of political opinion polling in the United Kingdom, from the first Gallup poll in Britain in October 1937 to the announcement of the new polling agency Deltapoll in March 2018. We note the history, techniques and personalities involved, gauge their effectiveness and speculate on future developments

## **English Housing Survey at 50: How is the EHS used by the Department for Business, Energy and Industrial Strategy to calculate fuel poverty statistics?**

Rebecca Cavanagh

*Department for Business, Energy and Industrial Strategy*

The department for Business, Energy and Industrial Strategy use the English Housing Survey to calculate fuel poverty statistics, which are published annually in a National Statistics report. The English Housing Survey is a national survey of people's housing circumstances and the condition and energy efficiency of homes in England. In 2017, the survey celebrated its 50th birthday. In 1967, the first year of the EHS, 2.5 million homes didn't have an inside WC. Now, we estimate over 2.5 million households are in fuel poverty. 50 years on, the EHS is still being used to monitor and evaluate the condition of the housing stock, and the experience of the nation's householders. The poster will summarise how BEIS use the EHS to calculate the fuel poverty measure and will draw out the key findings from our 2018 publication to explain: Who are the fuel poor? Who are most severely impacted by fuel poverty? It will focus on the dwelling and housing characteristics that affect fuel poverty, highlighting the equally important insights BEIS derive from both the EHS's physical survey and the interview.

## **Bivariate centiles from copulas and convex hulls**

Angie Wade, Mario Cortina Borja  
*University College London*

Population centiles are widely used to highlight individuals who have unusual values in the response variable and sometimes it is of interest to study more than one response jointly. It is often necessary to construct centiles adjusted for one or more covariates. Whilst the methodology for the construction of univariate centiles is well established, bivariate centiles cannot be uniquely defined. The benefits of bivariate centile modelling are particularly pronounced when there is a strong, non-uniform dependence between the response variables. However, constructing bivariate centiles is complicated as there are many regions containing a specified mass of the joint probability distribution. Copula models may be used to construct bivariate centiles, an application with sparse results in the literature. We follow an approach combining convex hulls and copula models to define such regions. A convex hull of a sample of bivariate points is a minimal set of points such that the line joining all these points is also part of the same set. They are polygons with all connecting lines forming internal angles of less than  $180^\circ$ , hence with no dents in their perimeter. Copulas are distribution functions of 2 or more dimensions that model the joint behaviour of response variables with known marginals. They can be used to model complex relationships between response variables that go beyond the bivariate normal distribution and incorporate a wide variety of marginal outcome distributions. In this poster, the joint relationship between Forced Expiratory Volume and Forced Vital Capacity is analysed to illustrate how convex hulls can act as the basis of distribution-free ways of exploring bivariate associations and how we can use them to produce parametric centiles with the assistance of copula models, given specific marginal distributions.

## Communicating Algebraic Statistics

Pasindu Perera, Hugo Maruri-Aguilar  
*Queen Mary University of London*

Techniques from Computational Commutative Algebra have recently been developed in the context of statistical analysis. This poster is based upon a summer project concerned with the study and application of recent algebraic techniques in a statistical problem. The project consisted of two parts. The first was a literature review of recent results for hypothesis testing in the context of contingency tables. In the second part of this project, some examples were developed with the aid of computer intensive code. Data for the examples was sourced partly from the European Social Survey (ESS), a biennial cross-national survey of attitudes and behaviour. The poster emphasizes the communication and teaching aspect of the problem, particularly intending to target potential student users.

References 1 Drton et al. (2009). Lectures on Algebraic Statistics. Oberwolfach Seminars. Birkhauser. 2 Fontana, Crucinio (2018). Markov Chain Monte Carlo sampling for conditional tests: A link between permutation tests and algebraic statistics. Preprint arXiv:1707.08513 3 Krampe, Kuhnt (2010). Model selection for contingency tables with algebraic statistics. In Gibilisco et al., Algebraic and Geometric Methods in Statistics. Cambridge University Press.

## **Predictors of Quality of Life of Stroke survivors & their informal caregivers evaluated at two tertiary care hospitals in Karachi-Pakistan**

Wardah Khalid<sup>1</sup>, Tazeen Saeed Ali<sup>1</sup>, Shafquat Rozi<sup>1</sup>, Michel T. Mullen<sup>2</sup>, Saleem Ilyas<sup>3</sup>, Ayeesha Kamal<sup>1</sup>

<sup>1</sup>*Aga Khan University, Karachi, Pakistan*, <sup>2</sup>*Perelman School of Medicine at the University of Pennsylvania, Philadelphia, USA*, <sup>3</sup>*DOW University of Health Sciences, Karachi-Pakistan*,

**Background:** Social and physical consequences of stroke are highly devastating and have profound influence on quality of life (QoL). The aim of study was to evaluate the impact of stroke on QoL and to identify important factors that affect QoL to inform actions that can improve the overall status of Pakistani stroke survivors.

**Methodology:** An analytical cross-sectional study was conducted. Stroke survivors underwent detail assessment encompassing their QoL scores, socio-demographic, post-stroke complications, stroke severity, psychological and functional disability by validated tools. Primary care givers were interviewed separately. Multiple linear regression technique was applied and  $\beta$  coefficients with 95% C.I were reported for important associations.

**Results:** The study was conducted at one private and one public tertiary care hospital in Karachi, the largest metropolitan city of Pakistan and home to all ethnicities. In total 350 dyads were recruited. Mean age of stroke patients was 57.14  $\pm$ 13.34 years and 68% were male, 63.71% lived in joint family system, 64% had good social support, 60% were depressed, 30% suffered severe disability and 70% post-stroke complications, only 11.71% had any sort of medical coverage and only 25.14% received post-stroke rehabilitation. Mean QoL scores were 164.18 $\pm$ 32.30. Caregivers were young, majority females, 51% changed their working hours to look after their loved one and 8% had to leave their job, 34% reported high stress whereas QoL scores showed higher scores. Multiple linear regression model depicted (Adj  $\beta$  95% C.I), that severe functional disability [adj $\beta$  -33.77(-52.44, -15.22)], depression [adj $\beta$ - 23.74(-30.61,-16.82)], hospital admissions [adj $\beta$ -5.51(-9.23,-1.92)] and severe neurologic pain [adj $\beta$  -12.41(-20.10,-4.77)] negatively impacted QoL of stroke survivors (P < 0.01).

**Conclusion:** The study showed that severe disability, depression, severe neurological pain, hospitals admission due to complications and dementia were the strongest predictors associated with QoL. Based on the important clinical factors cost effective interventions can be planned and tested in the future to improve lives of the survivor dyad.

## Understanding Immune Response using Statistical and Machine Learning Approaches

Jingjing Zhang,<sup>1</sup> Simone Cuff<sup>1</sup>, Ann Kift-Morgan<sup>1</sup>, Donald Fraser<sup>2</sup>, Nick Topley<sup>2</sup>, Matthias Eberl<sup>1</sup>

<sup>1</sup>Cardiff University, <sup>2</sup>Wales Kidney Research Unit; University Hospital of Wales, Cardiff University

Despite symptomatic variability in patients, unequivocal evidence that an individual's immune system distinguishes between different organisms and mounts an appropriate response is lacking. We here used a systematic approach to characterize responses to microbiologically well-defined infection in a total of 83 peritoneal dialysis patients on the day of presentation with acute peritonitis, based on a broad range of cellular and soluble biomarkers. The method of recursive feature elimination was adopted to find the best combination of biomarkers (Immune Fingerprints), based on which classifiers can distinguish between various pathogens and subgroups. Three classifiers were used: random forest (RF), support vector machines (SVMs) and artificial neural networks (ANNs). These three were compared in terms of their generalization performance. Correlation analysis and hierarchical clustering techniques were used to provide a detailed overview of the relationships between the biomarkers, whilst estimates of kurtosis and skewness were used to examine the distributions of the biomarkers, and ROC curves used to show each biomarker's ability to predict the outcome. The final fingerprints are confirmed based on these statistical analyses. The findings of this study have diagnostic and prognostic implications by informing patient management and treatment choice at the point of care. Our results demonstrate the power of machine learning tools when used to analyse complex biomedical datasets and highlight key pathways involved in pathogen-specific immune response.

## **Multilevel Modeling of Binary Outcomes in Complex Health Surveys: Assessing the Use of Methods of Clustering and Weighting**

Dr Shafquat Rozi<sup>1</sup>, Sadia Mahmud<sup>2</sup>, Gillian Lancaster<sup>3</sup>

<sup>1</sup>*Department Of Community Health Sciences, Aga Khan University, Karachi, Pakistan,*

<sup>2</sup>*Department of Medicine, Aga Khan University, Karachi, Pakistan,* <sup>3</sup>*Keele University,*

**Background:** It is important to understand the health seeking behavior of the population and trend of health services utilization in Pakistan. To investigate the determinants of health seeking behavior in Pakistan we suggest a multilevel pseudo maximum likelihood (MPML) approach to estimates model parameters for the complex survey design.

**Method:**The sampling strategy of the National Health Survey was stratified two stage cluster sampling. Overall 18,315 subjects were interviewed. This is three level data with PSUs at the third level, household at the second level and persons at the first level. Health care utilization was considered as a binary outcome.

**Results:** We found age, gender, marital status, household ownership of durable goods, urban/rural status, community development index, and province as significant predictors of health care utilization (p-value <0.05). We also found two significant interactions; between gender and marital status (p-value<0.005), and between the community development index and urban/rural status (p-value <0.045).The variances of the random intercepts are estimated as 0.135 for PSU level and 0.224 for households. The results are significantly different from zero (p-value<0.05) and indicate considerable heterogeneity in health care utilization with respect to HHs and PSUs. The estimated ICCs for household level and PSU level are 0.08 and 0.030 respectively.

**Conclusion:** Though we observed some divergence in the estimates of slope and variance between un-weighted and scaled weighted analysis but divergences are not marked. This may have occurred because of larger cluster size and relatively small ICC. Our study results reveal the inequalities between socio-economic groups and between urban and rural residents of Pakistan especially in terms of health care utilization. There is also a need to consider gender sensitive programs. This study gives advocates a stronger position in relation to decision makers in the government, as they marshal data to promote their policies for reform.

## **Distributed Big Data Processing for Economic Statistics at ONS**

Lewis Forder, George Zorinyants, Andy Banks  
*Office of National Statistics*

The Office of National Statistics (ONS) is the UK's recognised national statistics producer. A principal area of economic statistics is the analysis of trade in goods data, such as import and export figures for different countries and products. Recently, the ONS has pursued the development of techniques that enable a greater breakdown of trade in goods data to provide further granularity than has previously been possible. However, a principal challenge to increasing data granularity is the corresponding increase in dataset size, which presents issues for data processing (filtering, cleaning, linking datasets from different sources, applying adjustments, producing aggregates) and data storage (reliability and maintaining access permissions). To meet these challenges, the ONS has built and continues developing a Data Access Platform (DAP) utilising a Hadoop distributed file system, Hive databases for storage, and PySpark / SQL (Cloudera, Inc.) for calculations. We report two observations: (1) using DAP, processing that previously took 24 hours on a legacy non-distributed system can be achieved in under 30 minutes; (2) the processing of novel big data that comprises an 8-fold increase in storage size can be computed in around a few hours. Further, using DAP we have been able to apply improved methods for checking the statistical validity of data processes and to implement improved workflows for managing non-standard categories of trade data. In conclusion, a distributed storage and computing platform has increased the range of computational tasks that can be performed by the ONS, enabling a greater degree of data granularity, more efficient processing times, linking datasets from different sources and the implementation of improved statistical methods. Some of these granular data will be published in the UK national account's Blue Book 2018. On-going research is being carried out to find further aspects of data that can be provided to stakeholders in future publications.

## **Modeling the Timing of Aphids: Beyond the First Arrival Date**

Zhou Fang

*Biomathematics and Statistics Scotland*

The monitoring and prediction of insect populations is important in agriculture, with aphids in particular carrying a variety of plant diseases. In this case, suction traps may be used to produce measurements of insect numbers on a daily basis, over a long period of years, and we are interested in relating the timing of the appearance of aphids to environmental factors. Classically, attempts to model the timing of insects like aphids have centred on the first arrival date - that is, the first day that an aphid was caught in a trap in a given year. This value is however quite volatile, which raises questions about the appropriateness of modelling. Further, populations can vary greatly from year to year, which poses problems as years with relatively small numbers of insects become influential outliers in the dataset. We try to address this by modelling the entire distribution of aphid numbers over each year, assuming a parametric form with abundance, timing, and duration parameters. We then apply the LAML based MGCV methodology, facilitated by computational symbolic differentiation. This allows each of these parameters to have a nonlinear relationship to environmental variables with the potential for including simple random effects. This approach can also incorporate heteroscedastic or non-Gaussian distributions for the daily counts. Practical application of the new method show some promise in the appropriate and automatic handling of years with low abundances, though more work is in progress.

## **Simulation studies to quantify the impact of competing risk events in single-armed and comparative drug trials**

Prabin Dahal

*University of Oxford,*

Background: The Kaplan-Meier (K-M) is the currently recommended approach for deriving failure estimates for antimalarial drugs. Cumulative Incidence Function (CIF) which considers competing risk events (CRE) has been ignored in antimalarial literature. This work aimed to quantify the impact of CRE in single-armed and comparative antimalarial trials.

Methods: A clinical trial (n=500 patients) was simulated with 5, 10 and 15% documented primary endpoint (PE). For each scenario, different proportions of CRE were simulated to represent areas of increasing transmission (<10%, 10-20%, 20-40% and >40%). Time to PE and CRE were simulated using biologically plausible hazard functions. For comparative studies, nine scenarios which could be observed in a field-trial when comparing two drugs (drug A; drug B) were simulated. For each of these scenarios, log-rank test for comparing the equality of K-M curves and Gray's test for comparing the equality of the CIFs were used.

Results: In single-armed trial, the overestimation of cumulative failure of PE by the K-M method increased with increasing proportion of CRE. In high transmission areas, the maximum overestimation in failure was 0.75% at 5% PE and this rose to 3.1% and 4.3% when the drug failure was 10% and 15% respectively. In comparative trial, where drug B was associated with 2-fold increase in both PE and CRE (compared to drug A), the log-rank test appeared to be the more powerful test with a rejection probability of 99% compared to 90% with Gray's test. However, when drug B exerted differential effect on PE and CRE (i.e. reduced PE but increased CRE), the Gray's test was the more powerful of the two tests.

Results: In high transmission areas, where the risk of CRE is high, the competing risk analysis approach should be used for deriving failure estimates. For comparative trials, the choice of the test to establish difference should be guided by the research question of interest.

## Testing for differentially expressed genetic pathways with single-subject N-of-1 data in the presence of inter-gene correlation

Alfred Schissler<sup>1</sup>, Walter Piegorsch<sup>2</sup>, Yves Lussier<sup>2</sup>

<sup>1</sup>*University of Nevada, Reno*, <sup>2</sup>*University of Arizona*

Modern precision medicine increasingly relies on molecular data analytics, wherein development of interpretable single-subject ("N-of-1") signals is a challenging goal. A previously developed global framework, N-of-1-pathways, employs single-subject gene expression data to identify differentially expressed gene set pathways in an individual patient. Unfortunately, the limited amount of data within the single-subject, N-of-1 setting makes construction of suitable statistical inferences for identifying differentially expressed gene set pathways difficult, especially when non-trivial inter-gene correlation is present. We propose a method that exploits external information on gene expression correlations to cluster positively co-expressed genes within pathways, then assesses differential expression across the clusters within a pathway. A simulation study illustrates that the cluster-based approach exhibits satisfactory false-positive error control and reasonable power to detect differentially expressed gene set pathways. An example with a single N-of-1 patient's triple negative breast cancer data illustrates use of the methodology.

## **PPPI – The involvement of patients/people in the Design and Development of Clinical Trials within Ireland and Europe**

Jean Saunders,<sup>1</sup> Derick Mitchell<sup>2</sup>

<sup>1</sup>SCU/CSTAR@UL, *University of Limerick*, <sup>2</sup>IPPOSI

As a Statistical Consultant in Research Methodology and Design of all forms of medical research I have always had a particular interest in Patient Involvement and contributions towards the design of research studies. Clinical Trials were originally mainly designed and developed with input from various experts, statisticians, clinicians, nurses etc. Gradually there has been a recognition that maybe the patients (or participants) in the actual trial could contribute although at first this was mainly limited to their involvement in the choice of outcome measures. The difficulty of involving patients in clinical trial design was mainly attributed to patients not having sufficient understanding of the trials to be able to make more than simple design changes. However Patients' Organisations have developed within Ireland and Europe and have become more and more influential at all levels of treatment and health provision. Up until recently the patients/organisations still had very little input into clinical trial and other health research design leading to the patients and patients' organisations criticising the trials/studies when they were taking place or completed. This has led recently to the idea of the 'expert' patient – enthusiastically championed by Patients' Organisations. This is the concept that patients often know quite a lot about their disease so the only thing stopping them from providing useful input into a clinical trial is their comparative lack of knowledge of interpreting the medical literature and/or the design of efficient and valid clinical trials. To address this various programmes of education for patients have been devised in Ireland and within Europe training them how to 'read' medical papers with full critical appraisal and understand more comprehensively the methodology behind and results of clinical trials. This poster provides the history behind these initiatives and the (successful) results obtained. Now very few trials are planned in Ireland and the EU without some input from patients or patient groups.

## **A Novel Gene Expression Score accurately predicts biochemical recurrence of prostate cancer after radical prostatectomy**

Nahaa Alsubaie<sup>1</sup>, Dr Amar Ahmad<sup>2</sup>, Jacek Marzec<sup>3</sup>, Attila Lorincz<sup>2</sup>, Yong-Jie Lu<sup>3</sup>, Jack Cuzick<sup>2</sup>

<sup>1</sup>Centre for Cancer Prevention, <sup>2</sup>Wolfson Institute of Preventive Medicine, <sup>3</sup>Barts Cancer Institute, Barts and The London School of Medicine

Background: Prostate cancer(PCa) is the commonest cancer in men in the developed countries. Gleason score and PSA score are the most informative clinical predictors of PCa aggressiveness. However, additional genetic information is needed to improve disease stratification and subsequent management. Existing clinical tools have limited accuracy in detecting biochemical recurrence(BCR) in patients with localised PCa who are at risk of relapse. Though various genetic predictors have been published, few being commercially available for clinical use. We aimed to identify gene expression signatures that could improve the prediction of BCR after Radical Prostatectomy(RP).

Method: We used the publically available Cancer Genome Atlas (TCGA) PCa dataset to develop a novel gene expression signature that can improve the prediction of BCR in conjunction with clinical variables (e.g. Gleason and PSA). A novel clinical variable (CV) score was developed by fitting a multivariate Cox's proportional hazards model with Gleason score, log(1+PSA), age and age squared. The outcome was time to biochemical recurrence in 469 (n-BCR=90) PCa patients treated by RP. The median follow-up was 2.59 years (IQR: 4.07-1.45). The LASSO procedure was used to select the most significant genes from 45,198 genes. A novel gene expression(GE) score was developed with 11 selected top genes (with minimal cross-validation error).

Results: In univariate Cox model the interquartile hazard ratio (HR) of the GE-score was 9.238 (95% CI: 6.475-13.180, LR- $\chi^2$  p-value<2.2e-16), c-index was 0.883 (95% CI: 0.801-0.955). In a bivariate Cox model, the GE-score dominated (interquartile HR=8.148 (95% CI: 5.626-11.802), p-value<2.2e-16) with the CV-score  $\Delta\chi^2 = 144.1$ , p-value<2.2e-16). The c-index of the fitted bivariate model was 0.885 (95% CI: 0.812-0.957).

Conclusion: The derived gene expression score can be a useful tool that adds additional information to improve the prediction of BCR in patients with localized PCa. However, further validation studies are needed.

## **Analysis of elephant carcass locations in Etosha National Park using the new, and improved, CReSS with SALSA model selection.**

Lindesay Scott-Hayward, Monique Mackenzie  
*University of St Andrews*

This analysis is motivated by the MIKES dataset (Minimising the Illegal Killing of Elephants and other endangered Species; <https://cites.org/eng/prog/mike>) in Etosha National Park (ENP). The dataset comprises 320 carcass locations over 18 years from 2000 to 2017. We use this dataset to show the development of an improved selection method for regression models to replace the model averaging used in the original CReSS paper (Complex Region Spatial Smoother; Scott-Hayward et al 2014). We have enhanced SALSA 1D (Spatially Adaptive Local Smoothing Algorithm; Walker et al 2011) for use in a two-dimensional smoothing alongside the original CReSS basis function to allow the selection of knot number, location and effective range of each basis. Additionally, there is now an option for a Gaussian basis function alongside the existing exponential function and the existing choice of Geodesic (“as the fish swims”) or Euclidean (“as the crow flies”) distances. We present results of analyses using the old CReSS method and the new SALSA-based CReSS method using both basis types and both distance functions. All code to use the new method is readily available in the MRSea package in R ([github.com/lindesaysh](https://github.com/lindesaysh)). The methods were compared using 10-fold cross-validation. Not only is the new method much faster, computationally, but it is also able to identify surface features that we would otherwise have missed using the old method. For example, in a simple model containing only a two dimensional smooth of space, the new method identifies stripes where carcasses are found along roads. The old, model averaging method, smooths through these areas. References: Scott-Hayward, L.A.S., M. L. Mackenzie, C. R. Donovan, C. G. Walker & E. Ashe. 2014. Complex Region Spatial Smoother (CReSS). *Journal of Computational and Graphical Statistics*, 23:2, 340-360, Walker, C., Mackenzie, M., Donovan, C., and O’Sullivan, M. 2010. SALSA—A Spatially Adaptive Local Smoothing Algorithm. *Journal of Statistical Computation and Simulation*, 81, 179–191. 

## **How to adjust comorbidity in cancer prognosis and mortality outcome using claim-based healthcare big data: An application to liver cancer patients in Korea**

Sanghee Lee, Kyungeun Bae, Dahhay Lee, Hyunsoon Cho  
*National Cancer Center Korea*

**Background and Objective:** Cancer prognosis is frequently complicated by the presence of comorbid conditions, which limits treatment and leads to poor prognosis and outcomes. Prognostic prediction model lack of comorbidity risk adjustment would result in biased estimates, but there is no widely accepted algorithm for it. We therefore demonstrate various comorbidity measurement algorithms for cancer patients using claim-based healthcare big-data. And a refined measurement algorithm is applied to patients with liver cancer in Korea.

**Method:** Quan's ICD10- and Klabunde's rule out algorithms were used to identify Charlson comorbidity from the claim data. For each comorbid condition, prevalence and impact on mortality were compared according to look-back, window period, and inpatient/outpatient claim forms. We conducted survival analysis in the presence of competing risks where age, sex, and significant comorbid conditions were added. Prognostic performance of the model was evaluated using Harrel's c-statistic.

**Result:** Prevalence and beta-coefficient estimates from the survival model varies with comorbidity observation period and type of claim. In general, solely inpatient claims improved the performance of short-term mortality prediction. However, the inclusion of outpatient claims and longer look-back year were required to predict long-term mortality. In application to liver cancer, comorbidities were measured with 2 years look-back and 30 days window period using both in- and outpatient claims. Liver cancer patients with multiple comorbid conditions were at high risk for other-cause mortality. The Harrel's C-statistic were 0.71 in short-term, 0.74 in long-term mortality prediction.

**Conclusion:** We compare comorbidity measurement algorithms for risk adjustment in cancer prognosis and mortality. Estimates of risk-adjusted mortality are dependent on the comorbidity measurement algorithm, and the optimal algorithm is proposed based on the mortality outcome and prognostic performance.

This work was supported by National Cancer Center Korea, under grant No. NCC-1710300-2 and No. NCC-1710142-2, and National Research Foundation of Korea grant NRF-2016R1C1B1008810 funded by the Korea Ministry of Science, ICT and Future Planning

## **Development of A Natural History Model for the Analysis of Liver Cancer Occurrence in Hepatitis B-infected Patients in Korea**

Kyung Eun Bae, Sanghee Lee, Byung Woo Kim, Moran Ki, Hyunsoon Cho  
*National Cancer Center Korea*

**Background and Objective:** In Korea, liver cancer is the second leading cause of cancer related death (14.1% of all cancer, 2015), and hepatitis B has been reported as the most important risk factor of liver cancer (72.3% of all risk factors, 2010). However, there has been lack of studies on modeling of liver cancer occurrence in hepatitis B-infected patients. Therefore, we aimed to develop a natural history model of liver cancer occurrence in hepatitis B-infected patients in Korea using deterministic compartmental approach.

**Materials and Methods:** To develop a natural history model, we adopted deterministic compartment model with three disease states; chronic hepatitis B (CHB), liver cirrhosis (LC), and hepatocellular carcinoma (HCC). For estimation of the unknown model parameter, we used healthcare bigdata, the Korean National Health and Nutrition Examination Survey (KNHANES), as the objects of fitting by the model function. Using our model, we demonstrated how to predict disease prevention with control parameters, recovery rate of CHB and transplant rate of LC, and presented the results of their control effects. Additionally, we illustrated how to conduct uncertainty analysis for our model parameters using several statistical methods; Latin Hypercube Sampling (LHS) method and Partial Rank Correlation Coefficient (PRCC).

**Results:** Our model well predicted the degree of disease prevention by controlling for treatment-related model parameters, suggesting which disease is more affected by CHB recovery and which control is more effective on HCC prevention. And parameter uncertainty analysis reflected well model variability for the baseline curves.

**Conclusion:** In this study, we suggested model development process and its utilization for liver cancer occurrence in hepatitis B-infected patients in Korea. Acknowledgement

This work was supported by National Cancer Center of Korea Grant (grant No: 1710142-2).

## **Time Series Analysis and Forecasting of Demand for Electric Vehicles Using Auto Trader UK Search Data**

Jenny Burrow, David Hoyle, Agnes Altmets  
*Auto Trader UK*

Interest in and demand for electric and hybrid plug-in vehicles has increased rapidly over the past five years. Auto Trader is the UK's largest digital automotive marketplace with over 55 million cross platform visits each month. This gives us a large and unique dataset which can be analysed to gain insight into consumer interest in different vehicle types. Here we employ a variety of time series techniques to analyse the number of searches for electric vehicles on the Auto Trader UK website and apps. We explore the impact of changes in consumer preferences and regulatory announcements on searches for electric vehicles. In particular, change point detection algorithms are applied to determine whether there is any evidence that recent regulatory announcements have had a longer-term impact on the level of interest in electric vehicles. Our data also allow us to explore regional variation in the demand for electric vehicles, and compare this with the current locations of electric vehicle charging points, to highlight where charging infrastructure does not match the current interest levels.

## **The Optimal Group Size Controversy for Infectious Disease Testing: Much Ado About Nothing?**

Brianna Hitt,<sup>1</sup> Christopher Bilder<sup>1</sup>, Joshua Tebbs<sup>2</sup>, Christopher McMahan<sup>3</sup>

<sup>1</sup>*Department of Statistics, University of Nebraska-Lincoln*, <sup>2</sup>*Department of Statistics, University of South Carolina*, <sup>3</sup>*Department of Mathematical Sciences, Clemson University*

Group testing, the process of testing specimen amalgamations, is an indispensable tool for laboratories when testing high volumes of clinical specimens for infectious diseases. An important decision that needs to be made prior to its implementation involves determining what group sizes to use. In best practice, an objective function is chosen and then minimized to determine an optimal set of these group sizes, known as the optimal testing configuration (OTC). There are a few options for objective functions, and they differ based on how the expected number of tests, assay characteristics, and laboratory constraints are taken into account. These varied options have led to a recent controversy in the literature regarding which objective function is best. In our poster, we examine the most commonly proposed objective functions. We show that this controversy may be “much ado about nothing” because the OTCs, group sizes, and corresponding results (e.g., expected number of tests, accuracy measures) from using the two most commonly proposed objective functions are largely the same for standard testing algorithms in a wide variety of situations.

## **Methodology for the assessment of analytical techniques for Nuclear Forensics in the presence of missing data**

Hannah Jevans, Rhiannon Ellis, Roy Awbery

*AWE*

Nuclear Forensics is an essential component of the UK's nuclear security infrastructure, providing the capability for determining the origin of nuclear material found out of regulatory control. Nuclear Forensics draws upon analytical techniques to distinguish the characteristics of nuclear and radiological materials to narrow down their origin and authorised use. Nuclear Forensics makes use of historical databases, which creates several challenges. One of these is 'missing data', where certain characteristics have not been measured for all historical samples. Another is censored data, where only the fact that the measurement is below a certain value is recorded. These characteristics can be particularly problematic when using techniques such as machine learning, since the heterogeneity in the data can create artificial signatures and bias in the results. We present methodology that has been developed to assess the impact of these effects. This work has highlighted several issues with commonly used machine learning approaches to Nuclear Forensics that require further research so that they can be better understood.

## **Continuous density-based nonparametric distance scaling for spatio-temporal cluster analysis**

Antonia Gieschen, Jake Ansell, Belen Martin-Barragan, Raffaella Calabrese  
*University of Edinburgh*

Analysing data over both space and time is an issue in various areas of application including health, marketing and public services. Based on the spatio-temporal clustering algorithm ST-DBSCAN, we describe a method of clustering spatial time series while taking into account varying data point densities across space in a continuous manner via density-based distance weighting. The resulting clusters can not only inform decision-making through a deeper understanding of spatio-temporal data, but also be used for representative sampling of data and the generation of synthetic data sets. Our method is developed using data from National Health Service (NHS) Scotland Open Data on drug prescriptions. Possible applications reach further, e.g., for retailers and public services striving for an increased understanding of their customers while, at the same time, being concerned about retaining anonymity of identifiable single-person data. Our results demonstrate how, and offer a solution for, the necessity of methods adaptive to varying densities when performing spatio-temporal clustering of data points over large spatial areas. Further research is planned to develop an approach that allows for changes in the size of considered spatial areas ('zooming'), as well as for changes in cluster composition and memberships over time.

## **Sequential data assimilation of the 1D self-exciting process with application to urban crime data**

Naratip Santitissadeekorn

*University of Surrey,*

This presentation looks at developing a novel data assimilation method for a self-exciting crime. Existing commercial predictive policing software (such as PredPol) already in use in the field, grid the 2D domain to be policed by carrying out 1D maximum likelihood estimation in each grid box. This approach cannot easily take into account uncertainty quantification or sequentially update as new data arrives - both of which would be useful from a policing perspective. Unfortunately, standard sequential data assimilation (DA) techniques that would allow us to overcome this problem cannot be applied. Hence, we develop a new ensemble filter for the crime problem by taking the same philosophy from data assimilation. The filter can be applied to other crime-rate models. In order to test the effectiveness of the filter, we carry out both a synthetic and real data tests on the forecasts. The LA gang data is chosen as it is believed that excitation is a major driver for the attacks between gangs. This data raises common issues for crime prediction, such as the noisiness of the data and the Hawkes model is far from being a perfect model. However, we find a slight improvement in the probability forecasts using the filter.

## **Blood stasis therapy for traumatic injury: a prospective, single-arm, pre-post pilot study**

Mi Mi Ko, Soobin Jang, Jeeyoun Jung  
*Korea Institute of Oriental Medicine*

**Objectives:** Blood stasis is an important pathophysiologic concept in Traditional East Asia medicine. It has been considered to be a pathogenic factor in chronic and incurable conditions such as pain, infertility, cancer, coronary heart disease and others. The aim of this study was to investigate the effects of pain reduction by blood stasis treatment for blood stasis syndromes with traumatic injury.

**Methods:** A single-centre, prospective, pretest-posttest pilot study. The study included 73 patients with a trauma that occurred within the past 2 weeks who were admitted into Jaseng Hospital of Korean Medicine from August 2015 through December 2015. Of the 50 patients analyzed in this study (Mean age 33.52 yr, 42 female and 8 male). Triple Energizer-reinforcing Saam acupuncture set, herbal medicine (Dangkwisoo-san), and Wet cupping on tender point, which are Korean medical therapies generally performed to treat posttraumatic pain, were performed. The patients had to receive at least six sessions of treatment during the 2 weeks, with the subsequent treatment being performed within  $3 \pm 1$  days after the previous treatment. Numeric rating scale (NRS) score, general pain severity indicator was measured as a primary outcome measure. The blood stasis questionnaire, oximetry and patient's satisfaction were also measured.

**Results:** The mean size of subcutaneous bleeding (width, length) and NRS score as a general pain indicator significantly decreased over the visits (all  $p < 0.01$ , visit 1 vs. visit 6) And the patients showed significant improvements in the minimum and maximum value of peripheral perfusion index after the treatment ( $p = 0.011$ ,  $p = 0.15$ , respectively).

**Conclusions:** Our study reveals that blood stasis treatment for traumatic injury may help improve reduction of pain. The significant results observed in this study support some evidence of the theories of diagnosing Blood stasis pattern and treatments those pattern in Korean Medicine.

## **Impact of climate modes to temperature and precipitation extremes in Mexico**

Rebeca Perez-Figueroa, Rory Bingham

*University of Bristol*

The impact of large scale climate modes on temperature and precipitation extremes is analysed, specifically the Pacific Decadal Oscillation (PDO), El Niño-Southern Oscillation (ENSO) and the Tropical North Atlantic indices. Monthly and annual indices for temperature and precipitation extremes are obtained, linear trends and composites are used to describe their temporal variability as well as to assess the impact of the selected climate modes of variability. It was found that temperature and precipitation extreme indices are considerably affected by the large scale circulation patterns, displaying different regional patterns as a response to teleconnection patterns. The impact of El Niño Southern Oscillation is dominant across the country exhibiting in general opposite patterns between its warm and cold events. Likewise, temperature and precipitation extremes display an ENSO-like response regional pattern to the Tropical North Atlantic positive (negative) phases. The Pacific Decadal Oscillation has a dominant effect along the west of Mexico.

## **Model Averaging in a Multiplicative Heteroscedastic Model**

Alan Wan<sup>1</sup>, Xinyu Zhang<sup>2</sup>, Yanyuan Ma<sup>3</sup>

<sup>1</sup>*City University of Hong Kong*, <sup>2</sup>*Chinese Academy of Sciences*, <sup>3</sup>*Pennsylvania State University*

In recent years, the body of literature on frequentist model averaging in statistics has grown significantly. Most of this work focuses on models with different mean structures but leaves out the variance consideration. In this paper, we consider a regression model with multiplicative heteroscedasticity and develop a model averaging method that combines maximum likelihood estimators of unknown parameters in both the mean and variance functions of the model. Our weight choice criterion is based on a minimisation of a plug-in estimator of the model average estimator's squared prediction risk. We prove that the new estimator possesses an asymptotic optimality property. Our investigation of finite-sample performance by simulations demonstrates that the new estimator frequently exhibits very favourable properties compared to some existing heteroscedasticity-robust model average estimators. The model averaging method hedges against the selection of very bad models and serves as a remedy to variance function mis-specification, which often discourages practitioners from modeling heteroscedasticity altogether. The proposed model average estimator is applied to the analysis of two data sets on housing and economic growth.

## **The potential for vehicle safety standards to prevent deaths and injuries in Latin America**

Jonathan Kent

*TRL (Transport Research Laboratory)*

The potential for vehicle safety standards to prevent deaths and injuries in Latin America. The overall aim of this research study was to support the adoption of minimum vehicle safety regulations for vehicles globally, using the four Latin American countries of Argentina, Brazil, Chile and Mexico as a case study. This was done by performing a cost-benefit analysis of applying regulations in this region between 2020 and 2030. The method for the analysis consisted firstly of estimating the number of fatalities that would occur among car occupants, pedestrians and pedal cyclists in each of the four countries between 2020 and 2030, assuming current trends in vehicle safety continue. The second step was to estimate how many of these fatalities could be prevented by implementing each of the vehicle regulations. This reduction was converted to a casualty economic benefit, using the value of statistical life (VSL) approach. Thirdly, the cost of applying the regulations was estimated by predicting how many extra cars that would need to be fitted with each safety feature, and then multiplying this by estimates of the fitment cost of each technology. Finally, the estimated benefits and costs were combined to produce a benefit-to-cost ratio to assess whether or not it would be cost beneficial to implement the regulations, considering the impact on car occupants separately from the impact on vulnerable road users (pedestrians and pedal cyclists). The study concluded that if minimum vehicle safety regulations are adopted in Argentina, Brazil, Chile and Mexico in 2020, an estimated 24,500 lives could be saved from 2020-2030, generating a total economic benefit of \$12.3 billion USD. Furthermore, the overall benefit-to-cost ratios for both car occupants and vulnerable road users were found to be greater than 1, indicating that the benefits of implementing the regulations would exceed the costs. Therefore, it was recommended that each country should adopt the minimum standards.

## **Mental health and transport**

Sritika Chowdhury, Lauren Durrell  
*Transport Research Laboratory*

Mental health has received increased attention in the last few years, and is recognised as one of the primary causes of disability in the UK. The UK's Mental Health Taskforce notes that evidence demonstrates that improving outcomes for people with mental health problems supports them to achieve greater wellbeing, build resilience and independence and optimise life chances, as well as reducing premature mortality. There is, therefore, a critical need to understand the extent to which mental health can impact day-to-day life and necessary tasks such as travel. TRL, the UK's Transport Research Laboratory, conducted a study with the aim of identifying and understanding more clearly the nature of the relationship between mental health and transport choices. Both qualitative and quantitative methods were used to gather data from a random sample of around 400 people in the UK to examine the impact of mental health on mode choice. In addition to interviews, a stated preference survey was designed to explore the impact of mental health on the decision-making processes when choosing particular travel modes. The design of the survey encouraged participants to trade-off between different journey attributes (such as time, cost, potential delay, number of changes and level of crowding) in order to choose their preferred mode of transport between car, bus or train. Nested logit and multinomial logit models were used to analyse the data collected from the stated preference survey. Results from these models highlighted a number of differences in the importance of different journey attributes between four mental health groups (anxiety, depression, both and neither). There was some evidence suggesting inherent biases towards certain modes of transport. This study enabled us to establish that mental health plays an important role in travel choice and its role in influencing habits and perceptions associated with travel behaviour.

## **Use of assurance in oncology clinical trials with radiologically-assessed survival outcomes**

Masashi Shimura, Akira Fukushima, Tadashi Hirooka  
*Taiho Pharmaceutical Co .Ltd.*

Background: Sample sizes are conventionally derived using a power function conditional on fixed unknown parameters for a specified treatment effect. To avoid the need for a conditional on fixed parameters, some researchers have proposed “assurance” as an average power by evaluating the conditional power function across the range of parameters for the treatment effect. Furthermore, assurance has also been extended to clinical trials with survival outcomes such as overall survival and progression-free survival (PFS). However, the main focus centers around the uncertainty of the unknown treatment effect while radiological assessment interval is not discussed. In PFS, the exact progression date is unknown and assessment intervals may influence assurance. Therefore, changes in assurance via assessment intervals need to be investigated.

Methods: We calculated the assurance for oncology clinical trials with radiologically-assessed survival outcomes via simulation. Three interval scenarios ((1) every four weeks, (2) every six weeks, and (3) every eight weeks) were set. Assuming an exponential distribution for the survival outcomes and a total sample size of 85, we set three cases of the prior distribution for median survival time i.e., the mean of the prior distribution is lower than, equal to, or higher than the planned treatment effect.

Results and Discussion: The percentage decrease in assurance via assessment intervals was large when the prior mean was lower than the planned treatment effect. The decrease in Scenario (3) was three times lower than that in Scenario (1). Contrastingly, when the prior mean was high in Scenarios (1) and (2), the assessment interval effect was negligible.

Conclusion: Consideration of assessment intervals and treatment effect uncertainty is important when calculating assurance for clinical trials with radiologically-assessed survival outcomes.

## **Quant4Qual. A Case Study for Undergraduate Students Developing Confidence with Statistics**

Jack Winterton

*London School of Economics and Political Science*

Quant4Qual is a short empirical research methods course designed and taught by undergraduate students for fellow undergraduate students. The Quant4Qual project has explored the possibility of bringing together students from across academic departments in a space that has a (1) facilitatory teaching style (2) peer-to-peer learning (3) learning grounded in case-study analysis. The aim is to establish an accessible learning environment for students who have little or no prior backgrounds of the subject to develop an initial understanding of statistics. The course assumes no knowledge of statistics prior to participation, indeed, the only prerequisite is a desire to better understand an exciting area of social science research. Participants learn of some of the most common tools and techniques used to answer causal 'what if' questions. The course follows a series of 5 workshops lasting for 1.5hrs each. Drawing on qualitative feedback from participants and workshop facilitators, this poster outlines the key ways that the project has broken down barriers in communicating statistics to a non-expert audience. This poster sets out the three core values behind the development and delivery of the Quant4Qual project. Peer-to-Peer Learning Quant4Qual is unique in its ability to connect undergraduates from across all departments and all years of study. We aim to disrupt the idea that a student is either Quant or Qual when an appreciation of both is needed to learn about the insights offered by the social sciences. Facilitatory Teaching Style Quant4Qual offered those students wishing to pursue careers in academia the opportunity to develop their teaching skills. Importantly, the workshop facilitators have to communicate their understanding of statistics in a way that is accessible to a non-expert audience and respond effectively to the questions. Case study Analysis Throughout the course, there is a strong emphasis on interpreting findings from actual studies and discussing the broader social and philosophical implications of such studies.

## **Statistical Modelling of Road Traffic KSI Accidents and Casualties in Great Britain**

Mohammad Sheikh

*Kingston University London*

This research study is motivated by the United Nations Decade of Actions for Road Safety 2011-2020. In this quantitative research, secondary data from DfT-STATS19 Database run by UK Police and the Department for Transport (DfT) are used. A number of databases are constructed based on road traffic KSI 'accidents', associated 'casualties', involved 'vehicles/drivers', 'contributory factors' etc. A number of statistical techniques are applied to explore the accident data as well as to analyse and to model statistically. The independent variable(s) are analysed based univariate as well as multivariate techniques. The purpose of this research is to investigate the factors causing crashes and to develop statistical models to help in understanding how road accidents might be reduced based on 'zero-vision'. The study covers Great Britain (GB), although it will be generalised to the rest of United Kingdom and other developed countries. The feasibility of transferring knowledge on developing a fleet safety culture in GB and other developed countries will be investigated. Key Words: Road traffic KSI accidents, Associated Casualties, Involved Vehicles/Drivers, Contributory-Factors, Breath-Tests, Zero-Vision, DfT-STATS19 Database, Data-Manipulation, Statistical-Exploration, Statistical-Analysis, Statistical-Modelling, Poisson, Negative-Binomial, ZIP, ZINB, Lognormal, GLM.

## **Analysing and Explaining the Relationship between Crime Victimization and Fear of Crime between Individuals and Communities**

Bethany Ward, Andromachi Tseloni  
*Nottingham Trent University*

This presentation will outline the preliminary stages and propose the next steps of a PhD research project which aims to investigate the relationship between crime victimisation and fear of crime, both within individuals and between communities. Theories developing from social disorganisation theory, such as social capital and collective efficacy will be employed as an explanation of the spatial clustering of both crime and fear of crime within neighbourhoods of certain characteristics. The research also aims to take into account the effects of the crime drop on the relationships investigated, as it has been demonstrated that crime has become consistently more spatially concentrated over time (Ignatans and Pease 2016). The research will use a number of datasets, including: the Crime Survey for England and Wales (Secure Access and End-User-License); The Community Life Survey (CLS)(Special Licence Access) and; The Census. Additional sources of community level data may also be included, such as Points of Interest Data, and police recorded crime data. Modelling strategies to be employed are generalised linear modelling, multilevel multivariate modelling (MVML) and hierarchical structural equation modelling (SEM). MVML will be used to analyse the relationship between crime and fear of crime both within individuals and between communities, and to determine the proportion of their relationship explained by various individual and community level variables. SEM will then be used to assess the intricate relationships between crime, fear of crime, various sociodemographic and contextual variables, and social disorganisation related concepts. Predictions of crime victimisation will be created in the CLS, based upon the results of generalised linear modelling undertaken in the CSEW to allow for SEM to be undertaken. The poster will demonstrate how the use of a combination of data sources can allow for richer analysis, including multiple individual and contextual variables, but also how more research questions can be answered through using simple, and more innovative data merging techniques.

## Using pattern mixture models to adjust for non-ignorable missing data in longitudinal analyses – applied to PD clusters

Michael Lawton<sup>1</sup>, Yoav Ben-Shlomo<sup>1</sup>, Margaret May<sup>1</sup>, Fahd Baig<sup>2</sup>, Thomas Barber<sup>2</sup>, Donald Grosset<sup>3</sup>, Michele Hu<sup>2</sup>

<sup>1</sup>University of Bristol, <sup>2</sup>Oxford Parkinson Disease Centre, <sup>3</sup>Queen Elizabeth University Hospital, Glasgow

**Introduction:** Standard longitudinal analysis uses mixed effects random slope and intercept models (MEM) which are robust when data are missing at random (missing data is associated with observed data) but not when data is missing not at random (missing data is associated with unobserved data). Some models have been developed, such as pattern mixture models (PMM), which are more robust under such circumstances. We are interested in looking at the prognosis of data derived Parkinson's Disease subtypes in a prospective cohort study. Therefore, we compared MEM to PMM to determine whether our progression rate estimates in motor disability and cognitive impairment might be biased due to patients dropping out of the study.

**Methods:** We carried out a cluster analysis of 1,601 and 944 idiopathic PD patients, from Tracking Parkinson's and Discovery cohorts respectively. Four clusters were identified and we measured prognosis in motor disability (UPDRS part III) and cognition (MoCA) using MEM. We then repeated the analysis using PMM. For brevity we focus on the results from Discovery only.

**Results:** We had 18% of patients dropping out of Discovery so the potential for bias is present. Using MEM cluster 1 had the fastest motor progression at 2.8 (2.3-3.2) UPDRS III points per year and cluster 4 was the slowest at 1.6 (1.1 to 2.2). The progression rates were very similar repeating the analysis using PMM. Using MEM cluster 3 had the fastest cognitive decline at 0.27 (0.14 to 0.41) MoCA points per year and cluster 2 the slowest at 0.10 (-0.05 to 0.25). The progression rates were very similar repeating the analysis using PMM.

**Conclusions:** Repeating our analysis using PMM we found little difference when compared to our standard analysis. Hence, we can be relatively confident that drop-out has not biased any of our progression rate estimates.

## **A pilot study of whether fictional narratives are useful in teaching statistical concepts**

Andy Field, Jenny Terry  
*University of Sussex*

Children engage in story-based learning from a very early age, using metaphor to infer knowledge and description to create mental imagery (Egan 1988; Egan and Gillian 2016). Although qualitative data suggest that students benefit from story-based learning (Blackburn 2015) there is a dearth of tightly controlled experimental studies to demonstrate the efficacy of narrative-based teaching. This pilot study aimed to look at the feasibility and plausible effects of using a fictional narrative to teach 11 statistical concepts. Thirty-five (13 males and 22 females) participants aged 19-63 years ( $M = 31.35$ ,  $SD = 13.86$ ) were allocated randomly to statistical materials presented in a standard textbook, Socratic dialogue or fictional narrative format. All materials were adapted from a textbook that uses a fictional narrative (Field 2016). Participants were pre-tested for their knowledge of the 11 concepts, their maths anxiety, and state anxiety. After reading the materials, 22 multiple choice questions were used to assess their understanding of the statistical concepts. We also took Likert-scale measures of how engaging they found the reading materials. Participants engaged with the measures and materials, but there is work to be done on honing the reading materials. A linear model predicting the multiple choice scores (0-22) from dummy variables coding the reading conditions (narrative vs. textbook, Socratic vs textbook) was fit using Bayesian estimation with relatively broad priors. The 95% HPDI intervals indicated that the plausible effect of using fictional narratives compared to standard textbook presentation ranged from  $b = 0.30$  to  $4.91$  ( $M = 2.39$ ). With some further development it will be feasible to test the effect of using narratives to improve students' understanding of statistical concepts. Also, it seems that there may plausibly be a benefit to using narratives and that a larger scale study is warranted.

## Characterisation of CT Noise with Applications to 3D Printing

Sherman Ip,<sup>1</sup> Julia Brettschneider<sup>1</sup>, Thomas Nichols<sup>2</sup>

<sup>1</sup>University of Warwick, <sup>2</sup>University of Oxford

X-ray computed tomography (CT) can be used for defect detection in 3D printing. The object is scanned at multiple angles to reconstruct the object in 3D space. The process can be time consuming. The aim of this project was 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. X-ray 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 (aRTist) was used to simulate the scan and it was compared with the x-ray acquisition under the face of uncertainty. However 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. It adjust each pixel statistic so that inference was done by comparing each pixel with the majority of its local pixels, reducing the number of false positives.

*Winner of Best Poster at the 2018 Research Students Conference*

## **Using a discrete choice experiment of patient preference to inform the design and interpretation of clinical trials: a case study in osteoarthritis**

Bethan Copsey, B, Buchanan J, Dutton SJ, Fitzpatrick R, Lamb SE, Cook JA.  
*University of Oxford*

**Objectives:** The poster will present on how discrete choice experiments could inform clinical trial design and interpretation of the findings, using a case study on osteoarthritis.

**Methods:** A discrete choice experiment aims to explore which factors influence a respondent's choice between different alternatives. Participants with hip or knee osteoarthritis (n=300) completed 16 choice tasks, selecting which of 2 hypothetical medications they would prefer. Medications were described in terms of pain, function and stiffness, duration of treatment effect, and the risks of taking the medication.

To identify an optimal set of choice tasks (minimising the standard errors), coefficients from pilot data (n=20) were used as fixed prior parameter estimates. The results were analysed using mixed effects logistic regression. The model parameters indicate which of the characteristics are more important in patient decision-making.

**Results:** The results demonstrate the importance of different treatment characteristics for patients. For clinical trial design, the findings can inform the choice of primary outcome and target difference in the sample size calculation and the planned trial analysis.

The results could also inform the interpretation of trial findings, indicating whether the benefits of a treatment in terms of the primary outcome outweigh the risks from the patient perspective.

## **Statistical modelling for change detection in remote sensing time series: Current methods, applications, and pitfalls**

Katie Awty-Carroll,<sup>1</sup> Pete Bunting<sup>1</sup>, Andy Hardy<sup>1</sup>, Gemma Bell<sup>2</sup>

<sup>1</sup>*Aberystwyth University*, <sup>2</sup>*Environment Systems*

Land cover type contributes to global climate change through influence on the sequestration of carbon, the hydrological cycle, and the reflectance of heat. A change in land use therefore alters the surface properties of that land and how it interacts with the earth's processes. These changes can be abrupt (i.e. due to deforestation), gradual (such as decreasing land quality), or phenological (i.e. changes in the timing of growth and senescence of vegetation). In the past few decades, data from Earth observation satellites has become an integral part of understanding and monitoring global land use change. The opening of the Landsat archive in 2008 gave researchers free access to over 30 years of continuous satellite observations at a temporal resolution of 16 days. As a result, over the last decade several methods have emerged which attempt to model the complex and periodic nature of remote sensing time series at pixel level. By capturing the underlying seasonality, change can be detected as deviation from the fitted model and by examining trends. However, modelling such time series presents challenges. Careful processing is necessary to remove observations affected by clouds, snow, and sensor noise. As a result, observation frequency can be highly variable leading to poor model fits. The objective of this review was to assess the different approaches which have been taken in this area, including how and where those approaches have been applied, their limitations, and possible directions for future work. The results of this assessment suggest that many methods are limited because they do not incorporate all available data. Many methods also lack replicability and have only been applied to limited numbers of case studies. In general, modelling approaches still struggle to accurately capture variability in intra-annual cycles.

## **Longitudinal analysis of multi-site bone mineral density measurements**

Rachel Tribbick,<sup>1</sup> Frank Dondelinger<sup>1</sup>, Marwan Bukhari<sup>2</sup>, Jemma G. Kerns<sup>1</sup>, Peter Diggle<sup>1</sup>  
<sup>1</sup>Lancaster University, <sup>2</sup>University Hospitals of Morecambe Bay NHS

Bone Mineral Density (BMD) measurements are used alongside demographic data to identify osteoporosis and calculate fracture risk in NHS patients. Despite BMD measurements from both the lower spine and hip region being of medical interest, most dual X-ray absorptiometry (DEXA) scans only measure BMD in the non-dominant hip. There is therefore currently a lack of literature comparing BMD loss across skeletal regions. We have access to data from over 30,000 patients who have had their BMD measured at up to 12 locations, 4 in the lower spine and 4 in each hip, having been scanned at the Royal Lancaster Infirmary. In addition to BMD measurements, the dataset includes a range of medically relevant covariate data. The aim of this study was to compare BMD loss across regions in the lower spine and hips, as well as examining the relationships between BMD and covariate data at each site. We performed a marginal analysis of BMD at each site using a Random intercept and slope model (Laird & Ware, 1982) on the 7,000 patients with longitudinal data. This has allowed for comparison of covariate effects across the sites, including biologically interesting differences between male and female patients. These differences are presented in this study. An alternate longitudinal analysis of interest is based on a serial correlation model (Diggle, 1988). The results of this alternative model to the marginal BMD loss at each site are compared to the previous analysis. One advantage for the serial correlation model is its ability to capture non-linear trends, which in practice tends to become more important as the length of follow-up increases. Further, this alternative allows for extensions to higher-dimensional outputs. As such, a discussion of potential multivariate models for BMD loss which respect and reflect the underlying anatomy of the data concludes this study.

## **Dynamic approximate forecasting algorithms for count data with environmental application**

Ali Gargoum

*UAE University*

Summary: In this research we propose an algebraic quick approximate Bayesian algorithm for learning in complex high-dimensional processes. These processes change dynamically with the passage of time where observations are taken sequentially. In such dynamic processes which are described in terms of parametric models, the model parameters or states can summarize the information needed to forecast the future of the process. This means that the probability distributions of the state space are updated sequentially after observing the system at each time period. When the system is Gaussian (the states are normally distributed and the observations have Gaussian density), the posterior can be computed in closed form. However, when the system is not Gaussian, approximations are necessary. The approximate procedure that will be developed in this work is parallel the sequential updating in the normal case and is based on the Bayesian conjugate analysis. It is fast and more efficient in comparison to the simulation-based Monte Carlo Markov Chain (MCMC) methods. The algorithm is discussed with an environmental application for predicting the spread of gaseous waste after an accident is discussed. This can be used in prediction of environmental contamination, in the event of an accidental release of radioactive pollutants or chemical gases.

## **Shared frailty Modelling for Grouped Repairable Systems**

Bodunrin Brown, Matthew Revie, Stuart McIntyre  
*University of Strathclyde*

The need to account for heterogeneity in the analysis of multiple repairable systems data have been clearly outlined in repairable systems literature. A wide class of models have been developed to this effect in reliability literature to analyse repairable systems with various degrees of repair efficiency such as imperfect repair using Heterogeneous Trend Renewal Process, and minimal repair using Non- Homogeneous Poisson Process. A common approach of accounting for the unobserved variables is via frailty. Most of the existing studies, however, have focused on accounting for frailties at individual-systems level. Shared frailty by groups of systems have not been explored yet in literature. We considered a case where systems can be classified under different regions and each group of systems is exposed to varying degrees of risk which in turn is depicted by the varying failure patterns in each group. In the research, we proposed a Non-Homogeneous Poisson process with shared frailty model. In particular, the newly proposed shared frailty model used a log-normal distribution to characterize the unobserved factors of variation, while the Non-Homogeneous Poisson process model was able to account for the average failure pattern of the population. To estimate the model parameters given a series of observed system failure times, we made use of newton method to estimate the parameters. A simulation study and a real-world case study were conducted to demonstrate the developed methods.

## **A Bayesian inference approach for determining player abilities in football**

Gavin Whitaker<sup>1</sup>, Ricardo Silva<sup>1</sup>, Daniel Edwards<sup>2</sup>

<sup>1</sup>*University College London*, <sup>2</sup>*Stratagem Technologies*

We consider the task of determining a football player's ability for a given event type, for example, scoring a goal. We propose an interpretable Bayesian inference approach that centres on variational inference methods. We implement a Poisson model to capture occurrences of event types, from which we infer player abilities. Our approach also allows the visualisation of differences between players, for a specific ability, through the marginal posterior variational densities. We then use these inferred player abilities to capture a team's scoring rate (the rate at which they score goals) through a Bayesian hierarchical model. Finally we describe a Gaussian mixture model which captures the areas on the pitch a player has most influence for these abilities. We apply the resulting scheme to the English Premier League, capturing player abilities, before using output from the hierarchical model to predict whether over or under 2.5 goals will be scored in a given fixture or not. We also highlight the key areas where players have most impact for these abilities.

## **Application of cross-classified models in Individual Participant Data meta-analysis**

Polyxeni Dimitropoulou,<sup>1</sup> Rebecca Playle<sup>1</sup>, Mark Kelson<sup>2</sup>, Lori Quinn<sup>3</sup>, Monica Busse<sup>1</sup>  
<sup>1</sup>*CTR, University of Cardiff*, <sup>2</sup>*University of Exeter*, <sup>3</sup>*Teachers College, Columbia University, New York*

**Background:** Individual Participant Data meta-analysis (IPDMA) is increasingly used in the analysis of combined datasets. It offers greater power than traditional meta-analysis, alongside validation of the original dataset. It therefore lends itself to more complex and robust analytical procedures and subgroup analyses.

**Objective:** To conduct an IPDMA of five feasibility RCTs of differing exercise regimes in patients with Huntington's disease. Critically, participants at some sites participated in more than one trial.

**Methods/Models:** Two-level mixed cross-classified models (CCMs) were used to account for the cross-classification of studies and sites (where site and study were non-nested clusters). The models examined the effects of exercise on the Unified Huntington's Disease Rating Scale modified motor score (mMS), a measure of motor function, and adjusted for age, gender and baseline mMS. Exercise effects were allowed to vary across studies. ICCs for participants in the same site but different studies were calculated by dividing the site variance by the total variance, and likewise for the other types of ICC. A non CCM was also derived and included site as a covariate, not taking account of its combinations with Study. Traditional MA using aggregate data was also conducted.

**Results:** Results were similar between the CCM (intervention effect with 95%CI: 0.5 (-0.8 to 1.7), p: 0.472, site and study ICCs=0) and the non CCM (0.4 (-0.9 to 1.6), p: 0.575, study ICC=0). The traditional MA showed considerable heterogeneity (I<sup>2</sup>=68%).

**Conclusion:** Results obtained for the models with and without cross-classification were in agreement. Clustering was mostly apportioned to variation in outcome between individuals and studies, mostly due to difference in interventions (frequency, duration, intensity of exercise and participant response) and in the baseline HD severity in participants, rather than variation within site. Stronger clustering within sites would require the correct cross-classified model specification for accurate estimation of treatment effects.

## **Simple risk score to assign treatment for women at high risk for preterm preeclampsia**

Ulla Sovio, Gordon Smith  
*University of Cambridge*

**Objectives:** To derive a simple risk score for preterm preeclampsia in nulliparous women based solely on the maternal history model used in the ASPRE trial, and to compare its screening performance with 1) the original ASPRE algorithm using maternal history and 2) the current definition of high risk according to the NICE guidelines.

**Methods:** Data from the prospective Pregnancy Outcome Prediction Study (POPS) of nulliparous women was used (n=4,190). Participants were classified into high and low risk groups based on the NICE guidelines. Model coefficients from the ASPRE algorithm were translated into a risk score while preserving the relative weight of each coefficient. Logistic regression analysis to predict preterm preeclampsia and ROC curve analysis comparing the published algorithm and the simple risk score was performed. The score was dichotomised to top 10% and bottom 90% of predicted risk and the screening statistics were compared with those of the binary NICE definition.

**Results:** In the prediction of preterm preeclampsia, the area under the ROC curve (AUC) was lower only by 0.0077 compared to the AUC using the original algorithm. The absolute risk of preterm preeclampsia in the whole POPS was 0.7%. A risk score of  $\geq 30$  classified women into top 10% of risk and this was defined as screen positive. The risk ratio (RR) of preterm preeclampsia using this risk score cut-off was 13.2 (95% CI 6.3-27.7), sensitivity was 57.1% (37.5-74.8%), false positive rate (FPR) was 8.8% (8.0-9.7%), and LR+ was 6.5 (4.6-9.0). 11% of women screened positive using the NICE guideline which gave a higher FPR of 10.7% (9.8-11.7%).

**Conclusions:** A simple risk score to assess the need for aspirin treatment to prevent preeclampsia gives a lower FPR compared to the current assessment using NICE guidelines. This would mean ~2% reduction in the number of women treated with aspirin.

## Copula Based GAMLSS with Non-Random Sample Selection

Malgorzata Wojtys,<sup>1</sup> Giampiero Marra<sup>2</sup>, Rosalba Radice<sup>3</sup>

<sup>1</sup>*University of Plymouth*, <sup>2</sup>*University College London*, <sup>3</sup>*Birkbeck*,

Non-random sample selection is a commonplace amongst many empirical studies and it appears when an output variable of interest is available only for a restricted non-random sub-sample of data. In this presentation, we introduce a generalized additive model for location, scale and shape which accounts for non-random sample selection. The classical GAMLSS is extended by introducing an extra equation which models the selection process. Specifically, the selection and outcome equations are linked by a joint probability distribution which is expressed in terms of a copula. Moreover, we model the relationship between covariates and responses by using penalised regression splines, thus capturing possibly complex relationships. This approach allows for potentially any parametric distribution for the outcome variable, any parametric link function for the selection equation, several dependence structures between the equations through the use of copulae, and various types of covariate effects. We made the new developments available via the `gjrm()` function from the R package GJRM. We consider the study of the effects of insurance status and managed care on hospitalization spells. The association between admittance and length of stay may suggest the presence of specific selection mechanisms. Previous authors motivated the use of the gamma distribution to model the length of hospital stay, which led to the finding that model selection indeed existed for this data set. We consider a wider set of marginal outcome distributions, link functions and copulae. We also employ smooth functions of age and years of education. In this set-up the inverse Gaussian turned out to be the distribution most supported by the data. We also found that non-random sample selection is not present. Our result has important implications for the study of selection bias as it highlights the fact that using a more restrictive set of modelling choices may lead to unfounded speculations on the presence of certain selection mechanisms.

## **Sub-national population projections for Wales**

Joe Wilkes

*Welsh Government*

Population projections are important for planning the delivery of a range of public services. Examples include deciding how many doctors and nurses to train and how many houses and schools to build. They are also used to inform debate and policy making in areas such as environment, public health and the economy. There is a strong demand for sub-national population projections that can assist in planning at local authority level. Here, we present a mathematical model used by the Welsh Government to estimate the future population by age and gender for each local authority in Wales. The model uses trends for births, deaths and internal and international migration. The assumptions on which these trends are based are described in detail. Variants of the projections that use different fertility and mortality rates and different levels of migration are shown. Finally, the population projections that were made in past years using a similar methodology are compared to the mid-year population estimates. These estimates are produced annually by the Office for National Statistics using records of births, deaths and migration. Such comparisons can help quantify the success of the population projections.

## Univariate Bayesian Change-point Detection On Homogeneous Poisson Processes.

Anthonia Afuape

*Certara- Simcyp*

Daily effects such as the number of annual failures/ disasters in industrial facilities or in the number of annual cases of a particular diseases can be considered as a Poisson process. Therefore, developing an approach to predict a change occurring in such processes and their extent can be useful to science and to the society. Assuming an independent and identically distributed (iid) random variable  $Y_i$  represents the number of observations in individual years,  $i=1, \dots, n$ . A simple form of a change-point model can be described as  $Y_i \sim f(Y)$  from  $i=1, \dots, k$  and  $Y_i \sim g(Y)$  from  $i=k+1, \dots, n$ .  $k$  is the unknown parameter called the change-point and  $f(Y)$  and  $g(y)$  are known densities following a Poisson distribution with parameter  $\lambda_1$  and  $\lambda_2$ , respectively. When  $k=n$ , the model is interpreted to have no change. The objective of this study is to present the Bayesian estimation approach of parameters of Univariate Poisson change-point processes. For this, we defined a class of Prior distributions that possess a conjugate property and used it to obtain the Joint Posterior distribution. We assumed the Joint Posterior distribution is proportional to the product of the Likelihood function and known Prior distribution. Using Gibbs Sampler algorithm - an iterative Monte Carlo method, we were able to generate random samples to obtain characteristics such as mean and variance of our marginal posterior densities. Carrying out large enough iterations until the samples converged produced the mean estimates for our parameters of interest ( $\lambda_1$ ,  $\lambda_2$  and  $K$ ). We found our algorithm to be effective and produced accurate estimates for our parameters. In addition, we also noticed the algorithm to be sensitive to the number of observations ( $n$ ) and to the time the change-point ( $K$ ) occurs.

## **Bayesian deconvolution for Well Test Analysis**

Themistoklis Botsas, Jonathan Cumming, Ian Jermyn  
*Durham University*

Well test analysis is a set of methodologies used in petroleum engineering, which aims to use pressure and flow rate measurements in order to extract information about the wellbore and the reservoir. One of those methodologies is deconvolution, which is the process of inferring a response function from the data in order to obtain a description of the flow behaviour in the reservoir, and consequently gain insight into the system. We use two forms for the response function. One is based on key attributes of the well and reservoir that have been much used in the literature and in practice. The other is constructed as a combination of the exact solution of the diffusion equation for early times, and a 'theoretical' rectangular reservoir in association with the image method for late times. To make inferences about the response function, we use an errors-in-variables non-linear Bayesian regression likelihood. This allows us to account for the uncertainty in both the rate and the initial pressure measurements, which is essential in our context due to the large observational uncertainties encountered in practice. We combine the likelihood with a set of flexible priors for our parameters, which since the parameters are associated with the system, allows us to include information about their plausible form and range. We use an adaptive MCMC algorithm in order to approximate the posterior. We validate our algorithm by applying it to synthetic data sets. The results are comparable in quality to the state of the art solution, which is based on the total least squares method (in particular, we can model a wide variety of flow regimes), but our method has several advantages: we gain access to meaningful wellbore and reservoir parameters; we incorporate prior knowledge of the well and reservoir; and we can quantify parameter uncertainty in a principled way through the use of a Bayesian approach.

## Developing a Remote Statistical Monitoring Plan

Kirsty Wetherall, Heather Murray, Ian Ford  
*Robertson Centre for Biostatistics*

Background: Monitoring of clinical trials is necessary to ensure the protection of the study participants and the conduct of high-quality studies[1]. The aim of remote statistical monitoring is to carry out routine analyses of accumulating study data within a clinical trial to identify abnormal patterns, at individual study centres or groups of study centres, which might indicate deviations from the study protocol or regulatory guidelines. The expectation is that remote statistical monitoring will reduce the need for and the cost of on-site monitoring.

Methods: Remote Statistical Monitoring Plans (RSMP) should be developed at the beginning of the trial to specify the monitoring requirements. RSMPs should identify key data items to be monitored, frequency of reports, describe the statistical methods to be used to identify abnormal patterns of data and the processes for escalation of potential issues identified. No single approach to monitoring is appropriate or necessary for every trial therefore the RSMP should be adapted to each trial based on the data integrity risks of the trial and the consequences these could have on the safety of the study population<sup>1</sup>. As the trial progresses, it's likely the RSMP will have to be modified.

Conclusions: Potential issues that remote statistical monitoring may flag include under reporting of serious adverse events or endpoints, missing data, errors in data (outliers, reporting of incorrect units for laboratory data, miscalibration of instruments used in the collection of data), non-compliance to study protocol, possible fraud or lack of understanding of the protocol and delay in completing the Case Report Form (CRF) by on-site study staff. We will provide examples of output used in remote statistical monitoring reports and discuss potential advantages and disadvantages of remote statistical monitoring compared to traditional on-site monitoring.

*References: 1U.S. Department of Health and Human Services, Food and Drug Administration. Oversight of Clinical Investigations – A Risk-Based Approach to Monitoring. Available at [www.fda.gov/downloads/drugs/guidancecomplianceRegulatoryInformation/guidances/ucM269919.pdf](http://www.fda.gov/downloads/drugs/guidancecomplianceRegulatoryInformation/guidances/ucM269919.pdf)*

## **The Gamma Log-logistic Modified Weibull Distribution with Applications:**

Boikanyo Makubate,<sup>1</sup> Broderick Oluyede<sup>2</sup>

<sup>1</sup>*Botswana International University Of Science and Technology*, <sup>2</sup>*Georgia Southern University, Statesboro, GA, USA*

A distribution called the gamma log-logistic modified Weibull (GLLoGMW) distribution is presented. This distribution includes many submodels such as the log-logistic modified Rayleigh, log-logistic modified exponential, log-logistic Weibull, log-logistic Rayleigh, log-logistic exponential, log-logistic, Weibull, Rayleigh and exponential distributions as special cases. Structural properties of the distribution including the hazard function, reverse hazard function, quantile function, probability weighted moments, moments, conditional moments, mean deviations, Bonferroni and Lorenz curves, distribution of order statistics, L-moments and Renyi entropy are derived. Model parameters are estimated based on the method of maximum likelihood. Finally, real data examples are presented to illustrate the usefulness and applicability of the model.

## Covariate selection and doubly robust estimation for average treatment effects in high dimensional settings

Karla Diazordaz

*LSHTM,*

Electronic health records (EHR) are seen as valuable, cost-effective resources to answer questions about the effect of social and health interventions when randomisation is not feasible. However, because these databases are not collected for research, care needs to be given to controlling for confounding and dealing with missing data. Propensity score (PS) based methods have become increasingly popular in these settings to estimate the average treatment effect, which under certain conditions, can be interpreted causally. Building PSs involves selecting all confounding variables associated with the outcome and the exposure. However, this procedure becomes complicated in high-dimensional settings, common in EHRs. The use of variable selection strategies (or more general data-adaptive high-dimensional methods) is desirable, but may result in biased PS-based estimators. In contrast, provided certain assumptions hold, Double Robust estimators can be combined with variable selection and data-adaptive estimation, as their bias vanishes faster than the bias in either outcome or PS models (being of the order of the product of these two), achieving consistent estimates with valid confidence intervals. We describe two doubly robust (DR) estimators for valid inference after variable selection: (1) collaborative targeted minimum loss-based estimator (CTMLE) and (2) penalised bias-reduced double-robust estimation (PBRDR). A simulation study is used to compare standard Augmented Inverse probability of treatment weights (AIPTW), with two CTMLE versions and the PBRDR, in terms of the estimators' finite-sample performance (bias and coverage rate of the 95% confidence intervals), after variable selection and using data-adaptive fits for outcome and PS models, obtained using the Super Learner, an ensemble machine learning method. We also demonstrate the potential and practical utility of these methods by applying them to a large Brazilian EHR, to investigate the effect of conditional cash transfers on tuberculosis cure.

## **Generalized Meta-Analysis: A step towards building rich models by combining information from multiple studies**

Prosenjit Kundu, Nilanjan Chatterjee, Runlong Tang  
*The Johns Hopkins University*

In the world of decision making, data is an indispensable ingredient for answering relevant questions in almost all disciplines of study including science, humanities and business. Due to advancement in technology and ease in availability of modern tools, the research in a variety of fields including genomic medicine, genetics, clinical trials, epidemiology and environmental science has become data intensive with deluge of heterogeneous data. Our objective is to build rich models from disparate information available from these data sets. We propose developing a generalized meta-analysis (GMeta) approach for combining information on multivariate regression parameters across multiple different studies which have varying level of covariate information. Using algebraic relationships between regression parameters in different dimensions, we specify a set of moment equations for estimating parameters of a maximal model through information available from sets of parameter estimates from a series of reduced models available from the different studies. The specification of the equations requires a reference dataset to estimate the joint distribution of the covariates. We propose to solve these equations using the generalized method of moments approach, with the optimal weighting of the equations taking into account uncertainty associated with estimates of the parameters of the reduced models. We describe extensions of the iterated reweighted least square algorithm for fitting generalized linear regression models using the proposed framework. Based on the same moment equations, we also propose a diagnostic test for detecting violation of underlying model assumptions, such as those arising due to heterogeneity in the underlying study populations. Methods are illustrated using extensive simulation studies and a real data example involving the development of a breast cancer risk prediction model using disparate risk factor information from multiple studies. Further, the GMeta methodology is extended in a two-phase design—an efficient design in terms of cost, and the method is demonstrated using simulation studies from National Wilms' Tumor data.

## **A machine learning approach to the risk assessment of good primary and secondary schools**

Folasade Ariyibi

*Ofsted*

The risk assessment of good primary and secondary schools in England is currently used for assistance in scheduling short inspections. The current methodology, based on an unweighted scoring system, is used to create a RAG rating for each school, in relation to whether the short inspection is likely to result in a full section 5 inspection. This prediction is based on the school's likelihood to decline in inspection grade. The current methodology has the benefit of being transparent to users but a more complicated method could lead to better accuracy. It is therefore important to consider alternative methodologies that might provide better predictions as to whether a school is likely to decline in inspection grade after receiving a short inspection.

## **The Potential of PRKCB in the Promotion of Acetylcholine in Achieving Neuroplasticity to Increase Prognosis in Medulloblastoma**

Maria Fields

*miRCore*

Introduction: Medulloblastoma accounts for less than 2% of all brain cancers and is 18%-20% of pediatric brain cancers [1]. Out of the four subtypes, less is known about the pathogenesis of group 3 and group 4 subtypes. Group 3 medulloblastomas are generally metastatic and have the worst prognosis. Group 4 medulloblastomas frequently metastasize, but have intermediate prognosis compared to group 3 [2]. Comparing the expression of group 3 and 4 pediatric medulloblastomas may provide insight to finding potential genetic targets.

Methods: Dataset GSE37418 was downloaded from the NCBI public Gene Expression Omnibus database. 16 group 3 samples and 39 group 4 samples were analyzed in GEO2R. Then, the top 400 genes were studied in String db to identify kegg pathways and biological processes. Gene Ontology and Genecards provided information regarding the genes of interest and their biological processes. Kegg showed the involvement of genes within the chosen pathway.

Results: EGFR, FYN, PRKCB, MAPK8, MAPT are the selected genes of interests. EGFR, FYN, PRKCB, MAPK8, MAPT are upregulated in group 4. FYN and PRKCB are involved in the cholinergic pathway, which involves the transmission of acetylcholine. In the cholinergic pathway, PRKCB is indirectly linked to neuroplasticity. EGFR and MAPK8 promote cell proliferation. MAPT is associated with neurodegenerative disorders, like Alzheimer's disease.

Conclusion: The upregulation of PRKCB in group 4 poses as a promising target for improving the prognosis of the group 3 subtype. PRKCB effect on neuroplasticity in the cholinergic pathway may counter the proliferation and/or destruction caused by tumor cells resulting in better prognosis for group 4. Perhaps by amplifying PRKCB in group 3 patients, PRKCB would encourage neuroplasticity, ultimately increasing the acetylcholine level, improving prognosis.

References: 1. "Medulloblastoma." *American Brain Tumor Association*. n.d. Web. 16 April 2018. 2. "Pediatric Medulloblastoma – Update on Molecular Classification Driving Targeted Therapies." *NCBI*. 22 July 2014. Web. 16 April 2018.

## **Bayesian meta-analysis in medical evidence synthesis: a systematic review of design of methodology and its reporting**

Ram Bajpai, Josip Car

*Lee Kong Chian School of Medicine, Nanyang Technological University*

**Background:** Bayesian methods are becoming more popular in various areas of medical research, including meta-analysis. However, Bayesian meta-analysis (BMA) is not a preferred choice of evidence synthesis despite advances in computational methods, its appealing nature, and ability to overcome some of the difficulties encountered by the traditional methods.

**Objectives:** We present how BMA methodology has been implemented and reported in medical research over the years, to monitor the rate of adoption and provide an overview of the characteristics of published BMAs.

**Methods:** We searched review titles with 'Bayesian meta-analysis' phrase in Medline, Embase, CDSRs, CHINAL, PsycINFO via Ovid from its inception to December 2017. We identified 79 BMAs. We extracted data on the design and reporting of methodological quality. We performed a descriptive analysis for all the characteristics we extracted from the eligible BMAs.

**Results:** Two-thirds of (68.4%) BMAs were published after 2011. Primary author of these reviews was mainly affiliated to USA (31.6%), Canada (17.7%) and China (11.4%). Median number of studies and participants included in BMAs was 15.5 (range: 3-187) and 5103 (range: 206-6,528,639) respectively. Half of these studies (50.6%) used only randomised controlled trials followed by observational studies (27.9%), and mixed of both designs (19%). Fifty-seven percent of these studies justified their reason for choosing BMA over frequentist meta-analysis. Two-third studies (36.7%) did not report any information about choice of the prior distribution, 29.1% reported sensitivity analysis for the selection of appropriate prior distribution, 60% did not adequately define their simulation process, 48.1% did not report their choice of model (fixed- or random-effect), and 26.6% studies were mixed of frequentist meta-analysis and BMA. WinBUGS statistical software was the primary choice (45.6%) followed by the R (22.9%) for conducting BMA.

**Conclusions:** This review indicates suboptimal design and reporting of methodological quality. Standard reporting guidelines could facilitate better design and reporting of BMA.

## **Lessons to learn from the reporting of adverse events in randomised controlled trials: a systematic review of published reports in four high impact journals**

Rachel Phillips,<sup>1</sup> Victoria Cornelius<sup>1</sup>, Lorna Hazell<sup>2</sup>, Odile Sauzet<sup>3</sup>

<sup>1</sup>Imperial College London, <sup>2</sup>Drug Safety Research Unit, <sup>3</sup>Bielefeld University

**Introduction/Objective:** Randomised controlled trials (RCTs) provide an opportunity to compare rates of adverse events (AEs) between treatment arms allowing causality to be evaluated. However collection and reporting practices have been shown to be inadequate. We undertook a systematic review of journal articles to ascertain current approaches to the collection, selection, analysis and presentation of adverse events (AEs) in randomised controlled trials (RCTs). We identified examples of good practice and provide recommendations for future practice.

**Methods:** Original phase II-IV drug studies looking at the efficacy/effectiveness of an intervention published in the Lancet, BMJ, NEJM and JAMA from September 2015 to September 2016 were included. RCTs evaluating safety as the primary outcome were excluded. Using a standardised, pre-piloted, checklist we extracted data on trial characteristics, collection methods, assessment of severity and causality, reporting criteria, analysis methods and presentation of harm data.

**Results:** We identified 184 eligible trial reports (BMJ n=3; JAMA n=38, Lancet n=62; and NEJM n=81). Of which 62% reported on the methods used to collect AE information. AEs that cause patients to withdraw can be useful indicators of severity and impact to patients. Twenty-nine percent reported the number of withdrawals due to AEs and 21% included information on which AEs caused withdrawals. Results presented and analysis performed was predominantly on 'patients with at least 1 event' with 84% of studies providing no information on the number of events occurring. Despite a lack of power to undertake formal hypothesis testing, 46.7% reported p-values for binary outcomes. There was a pervasive practise (59% of studies) of categorising continuous clinical and laboratory outcomes.

**Conclusions:** Current reporting and analysis of AEs in trials is sub-optimal. Areas to improve include reducing information loss when analysing at patient level only and inappropriate practice of underpowered multiple hypothesis testing.

## Physical Activity in Breast Cancer Survival: a Meta-Analysis

Maria-Eleni Spej,<sup>1</sup> Vasiliki Benetou<sup>1</sup>, Evaggelia Samoli<sup>1</sup>, Francesca Bravi<sup>2</sup>, Carlo La Vecchia<sup>2</sup>, Christina Bamia<sup>1</sup>

<sup>1</sup>Dept. of Hygiene, Epidemiology and Medical Statistics, School of Medicine, National and Kapodistrian University of Athens, Greece, <sup>2</sup>Dept. of Clinical Sciences and Community Health, Università degli Studi di Milano, 20133 Milan, Italy

**Objective:** Physical activity may contribute to increased survival among women with a history of breast cancer, but the findings are still inconclusive. We have conducted a systematic review and meta-analysis in order to clarify the association of physical activity with overall mortality, breast cancer mortality and/or cancer recurrence among breast cancer survivors.

**Methods:** We searched PubMed database up to November 2017 for observational studies investigating physical activity in association to total mortality, breast cancer mortality and/or recurrence among adult women with previous breast cancer diagnosis. Physical activity was measured in MET-hours/week. Pooled hazard ratios (HRs) and 95% Confidence Intervals (CIs) were estimated with random-effects models. Heterogeneity was assessed with the estimate of between studies variability ( $T^2$ ) and  $I^2$  and quality assessment was evaluated through the Newcastle-Ottawa scale.

**Results:** Eight studies were included in the meta-analysis. During an average follow-up ranging from 3.5 years to 12.7 years there were 23041 participants, 1955 deaths from all causes, 739 deaths from breast cancer and 1398 recurrences/remissions. The average Newcastle-Ottawa score was 6.8 stars. Compared to women who reported low recreational physical activity (lowest quintile/quartile), women with high recreational physical activity levels (highest quintile/quartile) had a significantly lower risk of all-cause mortality (HR= 0.55, 95% CI 0.45-0.68) death from breast cancer (HR=0.62, 95% CI 0.42-0.93) and a lower, albeit not statistically significant, risk of recurrence (HR=0.81, 95% CI 0.56-1.16). However, there was evidence of heterogeneity across studies ( $T^2 = 0.0438$ ;  $I^2 = 52.4\%$ ).

**Conclusion:** Post-diagnosis recreational physical activity was associated with lower overall and breast cancer mortality. Breast cancer survivors may benefit from engaging in recreational physical activity. However, the role of bias, confounding and mainly reverse causation cannot be quantified from observational studies, particularly with reference to breast cancer specific mortality.

## **Oncology Phase II Adaptive Designs - Treatment effect estimates and their use in planning Phase III trials**

Arsénio Nhacolo, Werner Brannath  
*Universität Bremen*

New estimation methods for oncology Phase II adaptive designs We propose point and interval estimation for adaptive designs. We considered the recently proposed oncology Phase II two-stage single-arm adaptive designs with binary endpoint, in which the second stage sample size is a pre-defined function of the first stage's number of responses. Our approach is based on sample space orderings, from which we derive p-values, and point and interval estimates. Simulation studies show that our proposed methods perform better, in terms of bias and root mean square error, than the fixed-sample maximum likelihood estimator. Using Estimates from adaptive Phase II oncology trials to plan Phase III trials The clinical drug development is mainly done in three phases, Phase I, Phase II and Phase III. The knowledge gained in clinical trials of a particular phase is often used to plan trials of subsequent phases. That is the case with successful Phase II clinical trials in which, among others aspects, the effect size estimates are used to plan the sample size of the related Phase III trials. Due to small sample sizes, selection bias and other factors, Phase II estimates are often imprecise, resulting in inadequately powered Phase III trials. We evaluated through simulation studies the consequences, in terms of power, of using the effect estimate from Phase II adaptive design trials to plan sample size of Phase III trials in oncology. We used the naïve maximum likelihood and our proposed estimators for estimating the Phase II effect. Results showed that using naïve estimates lead to underpowered Phase III trials, while estimates that take into account the adaptiveness of the designs lead to power that is close to the target value.

## Improving the interpretation of randomised clinical trials in order to apply their results to individual patients

Huw Llewelyn

*Aberystwyth University*

The assumption of constant relative risk reduction for a randomized clinical trial (RCT) simplifies the calculation of absolute risk reductions for different baseline probabilities. However, it only provides an accurate approximation for low probabilities. The calculation is based on an underlying assumption that the effect of an intervention is to change the proportion with each outcome within the trial intervention limb. It also assumes that the likelihood distribution is the same for those with each outcome irrespective of whether the outcome was modified by an intervention. In order to provide valid probabilities between zero and one based on Bayes rule, it is necessary to assume a constant odds ratio (not a constant relative risk). In order to help readers of RCT reports, a graph can be displayed in the report to allow the probability of the outcome conditional on intervention to be read off from a curve by starting with the baseline probability of the outcome conditional on the control or placebo. The differences allow the 'number needed to treat for one to benefit' to be calculated for different values of the trial entry criterion (and not only for the single average of the values beyond the cut-off for the trial). This also models the way that an experienced doctor judges whether the patient's condition is mild and probably self limiting and less likely to benefit from treatment. The same mathematical model can be used to explore better predictors of outcome with and without intervention in accordance with the aims of 'precision medicine' and in order to reduce over-diagnosis and over-treatment[1]. Examples will be presented based on data from published randomized clinical trials.

*Reference: Llewelyn H, Ang AH, Lewis K, Abdullah A. (2014) Analyzing clinical trials to 'stratify' diagnostic and treatment criteria. In The Oxford Handbook of Clinical Diagnosis, 3rd edition. Oxford University Press, Oxford. p 633 -634.*

*<http://oxfordmedicine.com/view/10.1093/med/9780199679867.001.0001/med-9780199679867-chapter-13#med-9780199679867-chapter-13-div1-15>*

## **Confidence interval estimation of true prevalence for studies with small sample size and misclassification.**

Abin Thomas, Mohamed Hussein, Naila Shaheen

*King Abdullah International Medical Research Centre, Riyadh*

Estimating the prevalence of an outcome and its corresponding confidence interval is a recurrent problem in healthcare research. Several methods have been developed that have been shown to yield different coverage and precision at different sample size and population proportions. When the outcome of interest is subject to misclassification, estimating the true prevalence and the corresponding confidence interval requires additional adjustments. Current methods scale for misclassification by adjusting the upper and lower limits of the confidence interval by the sensitivity and specificity of the measure. These methods improve the coverage at the expense of precision. When sample size is small ( $n < 10$ ), the precision further deteriorate. Recently developed methods based on Edgeworth expansion has been proposed to improve precision [Zhou et al (2008)] near to zero and one. We extend this method by driving the Zhou-Li confidence limits to account for misclassification. Our method incorporates the sensitivity and specificity during the estimation of the cumulants. We assess the performance of our proposed method and compare it to other known methods using simulation studies. Simulation results suggest that our new proposed method performs better in terms of theoretical precision and coverage in studies with lower sample size. The coverage appears to vary according to the misclassification measurements in higher sample size even though the precision remains stable. So it is suggested that the method can be adopted for pilot studies which use screening tools for measuring the outcome.

## **Comparison Of Cox Proportional Hazard Model And Accelerated Failure Time Model With Application To Data On Tuberculosis/Hiv Patients In Nigeria**

Ogungbola Opeyemi Oyekola, Akomolafe Abayomi. A  
*Federal University of Technology. Akure*

Survival analysis has experienced remarkable growth during the latter half of the twentieth century. The methodological developments of survival analysis with profound influence are the Kaplan-Meier method for estimating the survival function, the log-rank test for comparing the equality of two or more survival distributions, and the Cox proportional hazards (PH) model for examining the covariate effects on the hazard function. The accelerated failure time (AFT) model was proposed but seldom used. In this thesis, we present the basic concepts, nonparametric methods (the Kaplan-Meier method and the log-rank test), semi-parametric methods (the Cox PH model, and Cox model with time-dependent covariates) and parametric methods (Parametric PH model and the AFT model) for analyzing survival data on Tuberculosis/HIV co-infected patients in Nigeria. We apply the methods to a cohort of these patients managed in tertiary Directly Observed Treatment Short Course (DOTS) centre, Nigerian Institute of Medical Research (NIMR) for the period of six months. Where we compare the effect of the accelerated failure time model with Cox proportional hazard model in determining the time to sputum conversion in TB patients who are co-infected with HIV. The research established that AFT model provides a better description of the dataset as compared with Cox PH model because it allows prediction of Hazard function, survival functions as well as time ratio. Moreover, PH model does not fit appropriately when compared with AFT model; thereby provide less appropriate description of survival data. The result revealed that the gamma model provided a better fit to the studied data than the Cox proportional hazards model. Hence, it is better for researchers of TB/HIV co-infection to consider AFT model even if the proportionality assumption of the Cox model is satisfied.

## **Constrained log-likelihood for partial proportional odds models**

Altea Lorenzo-Arribas,<sup>1</sup> Antony Overstall<sup>2</sup>, Mark Brewer<sup>3</sup>

<sup>1</sup>*Biomathematics and Statistics Scotland / University of Southampton*, <sup>2</sup>*University of Southampton*, <sup>3</sup>*Biomathematics and Statistics Scotland*

Partial proportional odds models are a flexible option to model ordinal response data which allows the proportional odds assumption to be relaxed for one or more covariates. However, they can be problematic, particularly when one of the covariates is continuous. In addition to issues of potential over-parameterisation and lack of convergence, they can predict negative class probabilities in special circumstances. We provide a simulation assessment of the frequency and reasons behind these problematic cases and propose two alternative solutions: firstly, by means of a Lasso penalisation; and secondly, through a re-parameterisation of the log-likelihood for the model. We compare the effectiveness of the proposed solutions via a case study looking at environmental attitudes.

## **Estimating a population cross tabulation from multiple data sources using the Generalised Structure Preserving Estimator (GSPREE)**

Kirsten Piller<sup>1</sup>, Joanna Taylor<sup>1</sup>, Alison Whitworth<sup>1</sup>, Angela Luna Hernandez<sup>2</sup>

<sup>1</sup>*Office for National Statistics*, <sup>2</sup>*University of Southampton*

Small Area Estimation is a suite of methodologies for generating estimates at fine spatial scales, for which survey data are either non-existent or too sparse to provide direct estimates of acceptable precision. ONS Methodology, in conjunction with the University of Southampton, have been researching Generalised Structure Preserving Estimation (GSPREE) as one method to produce small area population estimates for categorical population or household characteristics. GSPREE takes administrative data sources which contain information for the same set of areas and categories as the target population cross tabulation, but have issues such as they are outdated, have incomplete coverage or do not meet the target definitions. It supplements these sources with a social survey to model the small area estimates. GSPREE uses a log-linear model to relate the cross tabulations of the data sources, and benchmarks the estimated population cross tabulation to marginal totals. This poster describes how the GSPREE methodology has been used to produce ethnic group population estimates for all English Local Authorities by combining data from the outdated census, the Annual Population Survey and the School Census. Uncertainty in the GSPREE estimates are estimated using a bootstrap. The performance of the GSPREE estimator is also assessed in a validation scenario against the 2011 Census where the population distribution is known.

## On Spatial Models for Binary Data

Isabel Natario<sup>1</sup>, Paula Simões<sup>2</sup>

<sup>1</sup>*NOVA.ID.FCT, Quinta da Torre, Campus Universitário, 2829-516 Caparica, Portugal,*

<sup>2</sup>*CMA; Área Departamental de Matemática, ISEL - Instituto Superior de Engenharia de Lisboa, Instituto Politécnico de Lisboa*

Binary data are very common in public health applications, when modelling disease prevalence or incidence for example. Often data location is also an important factor to take into consideration, as in the example that illustrates this work regarding a national health line, where one is interested in modelling the factors that influence the probability of a line's user with an initial intention of going to an urgency room to change his/her mind about that, after calling the line. For those cases it is needed a spatial model for binary data. There are some alternatives based on logistic regression, as regressions using spatial covariates or generalized additive models with non-parametric effects on geographical coordinates of the data location, or logistic regressions with a built-in spatial function as in the inclusion of random effects or Gaussian process regression models for binary data. These are important when the purpose is to evaluate risk at an individual level, although some serious computational issues might exist, caused by fitting the spatial models to large datasets. Additionally, often for confidentially reasons or other, spatial information on individual level binary data is only available aggregated over some administrative regions, there is data misalignment. Because this information is crucial for the spatial model, it must be incorporated. There are several strategies that can be followed for doing that, from aggregating everything, the binary data, the covariate information and the spatial information, to only considering the latter aggregated. It is then important to understand what the impacts on the results of those different strategies are, which is pursued in this work using the above-mentioned health line dataset for illustration. Although results might be better for more disaggregated data, sometimes the benefits prove to be residual.

## Modelling Incident Reporting in the NHS

Chris Mainey<sup>1</sup>, Nick Freemantle<sup>2</sup>, Milena Falcaro<sup>2</sup>

<sup>1</sup>*UCL / University Hospitals Birmingham NHS Foundation Trust*, <sup>2</sup>*University College London*

Modelling Incident Reporting in the NHS  
The NHS incident reporting system (NRLS) is a national data warehouse of 'incidents' recorded in healthcare settings. The primary information in these reports is qualitative free-text descriptions of incidents, but the scale of the dataset (~1.8 million reports per year) makes national analysis impractical, as it is based on clinical staff reading individual reports. My PhD work has examined the use of statistical modelling techniques, and aggregation across different data sources, to build predictive models using NRLS. These models can be used by regulators and other organisations to drive statistical process control methods and identify organisations with systematically different reporting behaviours to aid learning and reduce harm. This poster presents the strength and weaknesses of these data, the process and rationale for aggregating 'exposure' data, and how 'noise' (including overdispersion and clustering) can be dealt with. The modelling types, methods, and comparisons of prediction error are presented. Models include Poisson GLMMs, Generalized Additive Models (GAMs) & regression trees (including 'boosting', 'bagging' and Random Forest extensions). Incident reporting models were well approximated by Poisson random-intercept models, without using compound distributions such as the Negative Binomial. Models were selected based on lowest mean absolute error (MAE) both in sample and applied to a subsequent period. Use of AIC was misleading due to residual overdispersion in the Poisson random-intercept model, smoothing parameter uncertainty of degrees of freedom in GAM, and no maximum likelihood estimate in tree models. This work illustrates a pragmatic approach to gaining insight from a 'noisy' dataset, 'borrowing' predictors from others sources. It shows how secondary use of data can add value, as data are already collected for another purpose. In this case, it will enabling NHS regulators to make better use of the data they already collect but are unable to tackle due to its scale.

## **An Application of the Gamma-Weibull Distribution to Raindrop Size Data**

Eno Akarawak, Olaide Abass, Benedict Iyere

*University of Lagos, Nigeria*

Rain drop size distribution (RDSD) is one of the most widely used phenomenon in studying rainfall. Different RDSD models namely exponential, lognormal, gamma and Weibull are been used in literature to study rain characteristics. In this research work, a recently introduced convoluted distribution, the Gamma-Weibull distribution (GWD), which has moments similar to those of the gamma distribution, is used to model raindrop size data and the fits were compared with that of gamma distribution. The raindrop size data were measured at four locations in Nigeria - Ife, Calabar, Enugu and Zaria. In order to model the data collected over a period of three years, GWD was re-parameterized to give it a rain drop size model form. The results of analyses show that GWD is competitive in fitting these data compared to the gamma model. It was therefore recommended that GWD be used as an alternative distribution to gamma in fitting rain drop size data.

## **A Closer Look at Non-Response: A Common Problem in Social and Business Surveys**

Alexandra Pop, Rhonda Hypolite  
*Office for National Statistics*

Our poster will address the issue of non-response in social and business surveys. We aim to define what non-response is, and how it is dealt with in each of these different types of surveys. We then want to focus on a particular business and social survey, and discuss how non-response affects results, and what the organization is doing to tackle this problem. Furthermore, we look at how ONS improves response rates in general, for both social and business surveys. With this, we want to leave our audience with a better understanding of the issue of non-response, and why it needs to be addressed.

---

**A**

Abushal, Tahani · 302  
Adams, Jeff · 148  
Adekunle Oyekunle, Kazeem · 291  
Ademola, Yasir · 314  
Afuape, Anthonia · 370  
Aitken, Colin · 201  
Aivaliotis, Georgios · 183  
Akarawak, Eno · 389  
Alegana, Victor · 283  
Alexander, Craig · 7  
Alih, Ekele · 310  
Allison, Katie · 6  
Alqahtani, Khaled · 320  
Al-shallawi, Ahmad · 318  
Alsubaie, Nahaa · 339  
Andersson, Per Gösta · 249  
Andreis, Federico · 220  
Ansell, Jonathan · 109  
Appel, Deirdre · 43  
Ariyibi, Folasade · 376  
Ashford, Chris · 74  
Attar Taylor, Eleanor · 52  
Awty-Carroll, Katie · 361

---

**B**

Bailey, R. A. · 38  
Baillie, Mark · 16, 295  
Bajpai, Ram · 378  
Balay, Iklm · 19  
Bampi, Vasiliki · 106  
Barnes, Darren · 4  
Beck, Pauline · 167  
Bell, Iain · 259  
Beran, Jan · 174  
Berckmoes, Ben · 108  
Bhatt, Samir · 281  
Bi, Jialin · 181  
Bland, Matthew · 230  
Bon, Joshua · 265  
Bond, Simon · 237  
Bonnett, Laura · 242  
Botsas, Themistoklis · 371  
Bowen, Isaac · 120  
Bowman, Adrian · 191  
Boyle, Laura · 215  
Bradley, Alexa · 200  
Bregantini, Daniele · 68  
Brettschneider, Julia · 306  
Brown, Bodunrin · 364  
Brown, Gary · 276

Browne, William · 152  
Bura, Efsthathia · 277  
Burn-Murdoch, John · 57  
Burrow, Jenny · 343  
Bwanakare, Second · 63

---

**C**

Caldwell, David · 75  
Carrington, Rachel · 308  
Castruccio, Stefano · 189  
Cavanagh, Rebecca · 328  
Cernat, Alexandru · 12  
Chang, Ya-Ting · 126  
Chatterjee, Nilanjan · 233  
Chigbu, Polycarp · 304  
Cho, Shu-Hsien · 20  
Chowdhury, Sritika · 352  
Cole, Tim · 86  
Collins, Gary · 30  
Columbu, Silvia · 173  
Copsey, Bethan · 360  
Cordell, Heather · 91  
Cornelius, Victoria · 105  
Cox, Sir David · 144  
Cuff, Simone · 315  
Curtice, Sir John · 284

---

**D**

Dahal, Prabin · 336  
Dai, Hongsheng · 153  
Darby, Sarah · 192  
Davies, Iyai · 297  
Davies, Jenny · 245  
Davies, Natasha · 45  
De Silva, Varuna · 9  
Dennis, Emily · 102  
Derrick, Ben · 70  
Di Caterina, Claudia · 72  
Diazordaz, Karla · 256, 374  
Didelez, Vanessa · 53  
Diggle, Peter · 221  
Dimitropoulou, Polyxeni · 366  
Ding, Peng · 254  
Dobson-McKittrick, Anselma · 23  
Donegan, Brendan · 246  
Du, Hailiang · 185  
Duguid Farrant, Trevor · 37

---

**E**

Ediriweera, Dileepa · 286  
Egan, Blaise · 40  
Ehrhardt, Beate · 169  
Eleni Spei, Maria · 380  
Eleuteri, Antonio · 21  
Ellis, Ciaran · 326  
Ensor, Joie · 1  
Eun Bae, Kyung · 342  
Evangelou, Evangelos · 188

---

**F**

Fairbairn, Rebecca · 33  
Fang, Zhou · 335  
Fauvernier, Mathieu · 48  
Ferguson, A. Nicole · 88  
Field, Andy · 177, 358  
Fields, Maria · 377  
Figueroa, Luciana · 305  
Flaxman, Seth · 263  
Flight, Laura · 162  
Folorunso, Serifat · 107  
Freni-Sterrantino, Anna · 240  
Fry, John · 252  
Fryer, Rob · 101

---

**G**

Gabry, Jonah · 190  
Galwey, Nicholas · 115  
García-Fiñana, Marta · 147  
Gargoum, Ali · 363  
Garrett, Greg · 275  
Ghosh, Sucharita · 209  
Gibbs, Chloe · 111  
Gieschen, Antonia · 346  
Goldstein, Harvey · 216  
Gordon, David · 172  
Granger, Emily · 194  
Grant, Charles · 156  
Greene, Karly · 247  
Gregory, Alastair · 279  
Gregson, John · 244, 325  
Griggs, Julia · 50  
Grilli, Leonardo · 35  
Groom, Gillian · 41  
Grubb, Anita · 110  
Gu, Fengyun · 307  
Guibourg, Clara · 25  
Guo, Liang · 182

Gusnanto, Arief · 243

---

**H**

Haneuse, Sebastien · 232  
Harris, Jenny · 121  
Hattle, Miriam · 59  
Hemming, Karla · 236  
Hicks, Mike · 116  
Higgins, Vanessa · 211  
Himali, Jayandra · 117  
Hitt, Brianna · 344  
Hogg, Cameron · 22  
Hooper, Richard · 235  
Hu, Shengwei · 176  
Huang, Chao · 154  
Huang, Jian · 309  
Huang, Xu · 303  
Hughes, David · 146  
Hughes, Rachael · 255  
Hui, Huaihai · 132  
Huitfeldt, Anders · 193  
Hutton, Jane · 203

---

**I**

Ip, Sherman · 359

---

**J**

Janikas, Mark · 187  
Jenkins, Jamie · 73  
Jevans, Hannah · 345  
Jiang, Luohua · 170  
Johnson, Kory · 96  
Johnson, Rob · 5  
Johnson, Toby · 90  
Johnston, Alison · 103  
Jones, Eilir · 204  
Jones, Rhian · 199  
Jostins-Dean, Luke · 92  
Jung, Tobias · 180

---

**K**

Kalina, Jan · 114  
Kanaan, Mona · 238  
Kartsonaki, Christiana · 131  
Kaye, Ella · 195  
Kejzar, Natasa · 3  
Kent, Jonathan · 351

Keogh, Ruth · 159, 224, 273  
Kim, Jihye · 171  
Kimber, Alan · 130  
Kin Hing Phoa, Frederick · 100  
King, Thomas · 178  
Kiwon, Francis · 119  
Knight, Keith · 208  
Ko, Mi Mi · 348  
Koneska, Elena · 124  
Konku, Adetola · 290  
Koskinen, Johan · 94  
Kosmidis, Ioannis · 62  
Krone, Tanja · 67  
Kuljus, Kristi · 15  
Kundu, Prosenjit · 375  
Kunst, Robert · 123

---

## **L**

Lancaster, Gillian · 125  
Latouche, Aurélien · 270  
Lawton, Michael · 357  
Leahy, Joy · 137  
Leckie, George · 151  
Lee, Sanghee · 341  
Leiby, Benjamin · 122  
Li, Bing · 278  
Lin, Xihong · 46  
Llewelyn, Huw · 17, 382  
Lorenzo-Arribas, Altea · 385  
Lowe, John · 87

---

## **M**

Ma, Chuoxin · 14  
MacKenzie, Gilbert · 217  
Madurasinghe, Vichithranie · 118  
Mainey, Chris · 388  
Makubate, Boikanyo · 373  
Marriott, Lorrae · 300  
Martin, Susan · 128  
Martyn Hill, Timothy · 251, 327  
Mathew Olayiwola, Olaniyi · 288  
Mayhew, Matthew · 166  
Mayo, Deborah · 143  
McCarthy, Omar · 42  
McCray, Gareth · 324  
McLernon, David · 160  
McManus, Sally · 51  
Mehrhoff, Jens · 78  
Meng, Xiangyu · 319  
Mercatanti, Andrea · 157  
Miller, Claire · 56

Moews, Ben · 8  
Morbey, Roger · 77  
Morey, Richard · 143

---

## **N**

Najeeb Albatineh, Ahmed · 292  
Najera, Hector · 248  
Natario, Isabel · 387  
Nelson, Fraser · 258  
Neocleous, Tereza · 313  
Ng, Kenyon · 61  
Nhacolo, Arsénio · 381  
Nicholls, Martin · 257  
Nickless, Alecia · 27  
Nikolaidis, Georgios · 138  
Noble, Alasdair · 298

---

## **O**

Olsen, Wendy · 93  
Oluwatosin, Oyetayo · 80  
Opeyemi Oyekola, Ogungbola · 384  
Osgood-Zimmerman, Aaron · 282  
O'Sullivan, Ian · 206  
Osuolale, Kazeem · 322

---

## **P**

Paccagnella, Omar · 36  
Pan, Yi · 155  
Panayotova, Plamena · 210  
Paris, Sandra von · 202  
Parker, Ben · 99  
Parker, Richard · 161  
Parry, Kevin · 76  
Patidar, Sandhya · 186  
Paul, Elieza · 321  
Pedder, Hugo · 60  
Perera, Pasindu · 330  
Perez-Figueroa, Rebeca · 349  
Perks, William · 127  
Perperoglou, Aris · 222  
Philips, Jack · 165  
Phillippo, David · 58  
Phillips, Rachel · 379  
Piller, Kirsten · 386  
Pina-Sánchez, Jose · 11  
Plewis, Ian · 274  
Ploubidis, George · 32  
Pohar Perme, Maja · 269  
Pop, Alexandra · 390

Popa, Aura · 312  
Popov, Valentin · 134  
Prattley, Jennifer · 10  
Prescott, Gordon · 129  
Prosdocimi, Ilaria · 82  
Putter, Hein · 271  
Python, Andre · 228

---

## Q

Quagliari, Anna · 135  
Quaresma, Manuela · 49  
Quill, Rachael · 268

---

## R

Rajballie, Aruna · 83  
Ramroth, Johanna · 196  
Riley, Richard · 28  
Robb, Matthew · 140  
Rodrigues, Eliane R · 25  
Rohrbeck, Christian · 84  
Rojas, Ilan Fridman · 24  
Rosati, Nicoletta · 112  
Roshini Sooriyarachchi, Marina · 289  
Rouanet, Anaïs · 226  
Rowland, Edward · 260  
Rozi, Dr Shafquat · 142, 333  
Rozi, Shafquat · 331

---

## S

Saegusa, Takumi · 231  
Saha, Saswati · 214  
Saleh, Asma · 323  
Salmaso, Luigi · 136  
Santitissadeekorn, Naratip · 347  
Sarychev, Andrei · 79  
Sauerbrei, Willi · 223  
Saunders, Jean · 338  
Schildcrout, Jonathan · 234  
Schissler, Alfred · 337  
Schneider, Deborah · 213  
Schneider, Ulrike · 97  
Scott-Hayward, Lindesay · 340  
Scruton, James · 272  
Sesia, Matteo · 175  
Seymour, Rowland · 163  
Sharkey, Paul · 267  
Sheikh, Mohammad · 355  
Shimura, Masashi · 353  
Smallman, Luke · 95

Smith, Dianna · 241  
Smith, Theresa · 239  
Snell, Kym · 29  
Song, Jiao · 141  
Sottosanti, Andrea · 299  
Sovio, Ulla · 367  
Spanos, Aris · 143  
Spencer, Neil · 18  
Spencer, Simon · 149  
Srakar, Andrej · 264, 287  
Stephenson, Victoria · 280  
Stockton, Phillip · 184  
Stolfi, Paola · 13  
Stubbings, Philip · 104  
Suprihatin, Bambang · 301  
Sutton, Andrew · 250  
Sweeting, Michael · 145

---

## T

Taiyari, KHADIJEH · 71  
Tavakoli, Shahin · 227  
Teece, Lucy · 2  
Telford, Alison · 168  
Thomas, Abin · 383  
Thomas, Marilyn · 64, 296  
Tickle, Samuel · 164  
Tinsley, Becky · 198  
Tiwari, Puneet · 133  
Toher, Deirdre · 179  
Tompsett, Daniel · 218  
Torkashvand, Elaheh · 139  
Tough, Fraser · 293  
Tribbick, Rachel · 362  
Truquet, Lionel · 158  
Trussart, Marie · 316  
Tsalamanis, Ioannis · 262  
Tucker, James · 225  
Turner, Heather · 207  
Turrell, Arthur · 261

---

## U

Utazi, Chigozie · 219

---

## V

Valberg, Morten · 69  
Vansteelandt, Stijn · 54  
Vidotto, Davide · 34

---

**W**

Wade, Angie · 329  
Waite, Timothy · 98  
Wan, Alan · 350  
Ward, Bethany · 356  
Washbrook, Liz · 31  
Watkins, William · 89  
Watt, Hilary · 212  
Weir, Ruth · 229  
Wells, Claudia · 44  
Wesonga, Ronald · 317  
Wetherall, Kirsty · 372  
Whitaker, Gavin · 365  
Wilkes, Joe · 369  
Wille, David · 113  
Wilson, Amy · 66  
Winterton, Jack · 354  
Wisniowski, Arkadiusz · 253, 294  
Wojtys, Malgorzata · 368  
Wolf, Levi · 150

Wood, Simon · 47  
Woods, David · 39  
Wright, Neil · 26

---

**X**

Xiu, Zhaoyan · 311

---

**Y**

Yates, Megan · 205  
Young, Grace · 197

---

**Z**

Zaman, Qamar · 81  
Zhang, Jingjing · 332  
Zidek, James · 266  
Zorinyants, George · 65, 334

