MEDICAL RESEARCH COUNCIL
CONFERENCE ON BIOSTATISTICS

In celebration of the MRC Biostatistics Unit's Centenary Year

Poster Session

Tuesday 25 March 2014

Quadratic discriminant analysis in longitudinal studies

Authors:
Riham El Saeti (presenting)
Department of Biostatistics, University of Liverpool
Riham.El-Saeiti@liverpool.ac.uk

Cheyne, Christopher
Department of Biostatistics, University of Liverpool

Czanner, Gabriela
Department of Biostatistics and Department of Eye and Vision Science, University of Liverpool

Poster Session **17.15-19.00** **Old Hall and Kitchens**

Keywords: covariance structure
longitudinal data
misclassification error
discriminant functional analysis

Abstract:
Linear discriminant function analysis is often used to classify individuals into two or more groups. We propose a discriminant analysis approach when both longitudinal information (measurements taken over time) and covariate information (such as age, gender, etc.) are involved in the same model. One of the challenges is to construct appropriate covariance matrices that accounts for the correlations between measurements over time and cross-sectional covariates. Also, when the classification relies on a longitudinal outcome, the underlying assumption of homogeneity of the variance-covariance structure rarely holds. Our approach will be applied to a cohort of patients with neovascular age-related macular degeneration.

Medical Research Council Conference on Biostatistics
in celebration of the MRC Biostatistics Unit's Centenary Year

24th - 26th March 2014 | Queens' College Cambridge, UK

Missing data and survival analysis of central nervous system tumours amongst children and young people in Yorkshire, 1990-2009

Authors:
van Laar, Marlous (presenting)
University of Leeds
m.vanlaar@leeds.ac.uk

Greenwood, Darren
University of Leeds

Feltbower, Richard
University of Leeds

Poster Session **17.15-19.00** **Old Hall and Kitchens**

Keywords:
cancer
multiple imputation
survival
CNS tumours

Abstract:
Background: Missing data is a common problem in medical research; in particular the severity of cancer at diagnosis is poorly recorded, therefore this prognostic variable is sometimes excluded from analysis. We investigated survival trends of central nervous system (CNS) tumours in children and young adults whilst using multiple imputation (MI) to impute missing values of disease severity (grade of tumour) and ethnicity.
Materials_and_Methods: Children and young people (<30 years) diagnosed with CNS tumours were identified from a population based cancer register in Yorkshire. Missing values for grade (I-IV) and ethnicity (White, Asian, Other) were imputed using ordinal and multinomial logistic regression respectively, with age, sex, year of diagnosis, deprivation, relapse and treatment as predictors. We performed 40 imputations, and assumed the data were missing at random. After MI, pooled hazard ratios (HR) were obtained via Cox regression models and results compared to a complete case analysis (CCA).
Results: A total of 795 cases met the inclusion criteria. Overall, missing data of one or both grade and ethnicity variables occurred in 30% of cases. After MI, survival analysis showed an increased risk of death for 'other' compared to 'white' ethnicity (HR=2.1; P = 0.034), and an increased risk of death for those with grade II, III and IV tumours compared to grade I (HR=3.5, 6.4 and 10.4 respectively). Additionally, survival improved significantly by 4% per year over the study period (P=0.001). There was no difference in survival by ethnic group or over time in the CCA, however, effects of grade were similar. MI reduced standard errors of coefficients by an average of 18% when compared to CCA.
Conclusion: MI was used to minimise bias and enhance the precision of analyses, offering advantages over CCA. Survival rates varied by grade and ethnicity, and showed a significant improvement over time.

Medical Research Council Conference on Biostatistics
in celebration of the MRC Biostatistics Unit's Centenary Year

24th - 26th March 2014 | Queens' College Cambridge, UK

Bayesian data analysis of genomic microarrays for detecting Streptococcus pneumoniae multiple serotype carriage

Authors:
Newton, Richard (presenting)
MRC Biostatistics Unit, Cambridge
richard.newton@mrc-bsu.cam.uk

Hinds, Jason
St. George's London

Wernisch, Lorenz
MRC Biostatistics Unit, Cambridge

Poster Session **17.15-19.00** **Old Hall and Kitchens**

Keywords:
Bayesian
serotying
streptococcus pneumnie
microarray
multiple carriage

Abstract:
Genomic microarrays can be used to test clinical samples for the presence of pathogenic organisms. The method is capable of providing more detailed information than established tests, including information on the presence of mixtures of species or strains and their relative abundance. Here we investigate the data analysis problems presented by such arrays using a serotyping microarray for Streptococcus pneumoniae as an example.

The technique presents a number of data analysis challenges. Noise is a problem, particularly at low relative abundance, as is cross-hybridisation. In the example discussed here there was a requirement to distinguish between 91 strains (serotypes) of Streptococcus pneumoniae using a microarray containing probes for 432 capsular genes. Each strain of the bacterium contains a small subset of these 432 genes, ranging in number from 1 to 22. However many of the strains have very similar combinations of capsular genes, making it difficult to distinguish the identity of a strain in a sample, particularly when the sample contains a mixture of strains. In addition some pairs of strains have identical complements of capsular genes necessitating extra probes on the microarray which have to be incorporated into the analysis.

We present statistical solutions to these analysis problems based on Bayesian methods. The Bayesian approach enabled the development of a flexible and expandable statistical model, which produced a robust and highly accurate analysis of the data.

Medical Research Council Conference on Biostatistics
in celebration of the MRC Biostatistics Unit's Centenary Year

24th - 26th March 2014 | Queens' College Cambridge, UK

Assessing vaccine effectiveness using observational data in the presence of hidden confounders

Authors:
Rodgers, Lauren (presenting)
University of Exeter Medical School
l.r.rodgers@exeter.ac.uk

Lin, Nan
University of Exeter Medical School

Henley, William
University of Exeter Medical School

Poster Session **17.15-19.00** **Old Hall and Kitchens**

Keywords:
observational study
hidden confounding
bias
prior event rate ratio adjustment
vaccine effectiveness

Abstract:
Primary care electronic databases, such as the Clinical Practice Research Datalink, provide a potentially rich source of information for assessing the effectiveness of interventions in clinical practice. However, when making such evaluations in the absence of randomisation, it is important to be alert to the risk of bias from unmeasured confounding. A motivating example for this study is estimation of theN effectiveness of the influenza vaccine in different 'at risk' population groups.N We investigate prior event rate ratio adjustment and other quasi--experimental analytic methods for assessing vaccine effectiveness using observational routinely collected data. The methods are based on the assumption that differences in outcomes between vaccinated and unvaccinated (control) groups before the vaccination period reflect the combined effect of all identified and unidentified confounders related to that outcome.
The proposed approach is explored using a simulation study with settings chosen to replicate key features of data on antibiotic prescriptions in elderly patients before and after influenza vaccination. We examine the influence of unmeasured binary and continuous confounders, which may be time dependent and unbalanced between vaccinated and control subjects. We allow for all-cause mortality as a competing risk in the simulation models. Attention is also given to the scenario in which vaccination has no effect for a subgroup of patients, of relevance to vaccine efficacy studies in ageing populations. The quasi--experimental methods are applied to data from the Clinical Practice Research Datalink and comparisons made with estimates of vaccine effectiveness from conventional models.

Medical Research Council Conference on Biostatistics
in celebration of the MRC Biostatistics Unit's Centenary Year

24th - 26th March 2014 | Queens' College Cambridge, UK

Spatial Regression and Spillover Effects in Cluster Randomised Trials with Count Outcomes

Authors:
Alexander, Neal (presenting)
London School of Hygiene and Tropical Medicine
neal.alexander@ishtm.ac.uk

Anaya-Izquierdo, Karmin
University of Bath

Lenhart, Audrey
Centres for Disease Control, Atlanta USA

Poster Session **17.15-19.00** **Old Hall and Kitchens**

Keywords:
randomised trials
cluster randomised trials
spatial analysis
indirect effect
spillover effect

Abstract:
Standard analyses of cluster randomised trials (CRTs) assume the outcomes from different clusters are independent. CRTs against infectious diseases are commonly carried out in geographically defined clusters which potentially induces betweencluster spatial correlation and indirect effects. We describe a methodology to analyse data from CRTs with count outcomes, taking such effects into account. We use spatial regression models with Gaussian random effects, where the individual outcomes have marginal distributions overdispersed with respect to the Poisson and the corresponding intervention effects have a marginal interpretation. These random effects model spatial dependence using a homoscedastic modification of the intrinsic conditional autoregression (ICAR) model, and the indirect effects are modelled using the distance to, and a novel measure of depth within, the intervention arm. We illustrate the methodology using data from a pair-matched CRT against the dengue mosquito vector Aedes aegypti, done in Trujillo, Venezuela.

Medical Research Council Conference on Biostatistics
in celebration of the MRC Biostatistics Unit's Centenary Year

24th - 26th March 2014 | Queens' College Cambridge, UK

Using the package "TestSurvRec" in R-language to compare survival curves with recurrent events

Authors:
Martinez, Carlos (presenting)
University of Carabobo, Bolivarian Republic of Venezuela
cmartin@uc.edu.ve

Poster Session **17.15-19.00** **Old Hall and Kitchens**

Keywords:
Statistical Tests
Recurrent Events
R-language
Survival Curves
TestSurvRec

Abstract:
The "TestSurvRec" package implements statistical tests to compare two survival curves with recurrent events. The survival curves of the groups are estimated using *Peña-Strawderman-Hollander* estimator. The package can be used to plots recurrent event data useful to survival analysis. The package contains two data sets; one data set refers to Byar's experiment that contains the recurrence of bladder cancer tumours in patients treated with pyridoxine and thiotepa and one third group considered as placebo. The other data set refers to the rehospitalisation times after surgery in patients with colorectal cancer. The study took place in the Hospital de Bellvitge, a 960-bed public University hospital in Barcelona (Spain). Four hundred and three patients with colon and rectum cancer have been included in the study. The variables considered were: *sex, age (<60, 60-74, 75 years), tumour site (rectum, colon), tumour stage (Dukes classification: A-B, C, or D), type of treatment (chemotherapy, radiotherapy), distance from living place to hospital (30 km, >30 km.), educational level (less than primary, primary, secondary, university)*. Both data sets are used to explain as works the implemented functions on the package "TestSurvRec". Recurrent events are common in many areas: *psychology, engineer, medicine, physic, biology, economics* and so on. Such events are very common in the real world, examples: *viral diseases, carcinogenic tumours, machinery and equipment failures, rains, industrial accidents a*nd so on. Survival analysis is a branch of the statistic that it is used to model the time until the occurrence of events. Its objectives are: modelling of the survival functions, estimations of the risk functions of the occurrence of an event, estimations of probabilities of occurrence and the comparisons of survival curves of population groups. With this package, we can make these comparisons. The "TestSurvRec" package is available on CRAN-R.

Medical Research Council Conference on Biostatistics
in celebration of the MRC Biostatistics Unit's Centenary Year

24th - 26th March 2014 | Queens' College Cambridge, UK

Analysis of longitudinal outcomes allowing for complex dropout processes

Authors:
Kolamunnage-Dona, Ruwanthi (presenting)
Department of Biostatistics, Institute of Translational Medicine, University of Liverpool
kdrr@liv.ac.uk

Williamson, Paula
Department of Biostatistics, Institute of Translational Medicine, University of Liverpool

Powell, Colin
Department of Child Health, Institute of Molecular and Experimental Medicine, Cardiff University School of Medicine

Poster Session **17.15-19.00** **Old Hall and Kitchens**

Keywords:
Competing risks
Dropout
Joint modelling
Longitudinal data

Abstract:
Methods for the combined analysis of longitudinal and survival data have been developed considerably in the past decade for analysing sequential longitudinal measurements on each individual when sequence can be terminated early through withdrawal from the study (dropout). However, simple approaches such as complete case analysis still remain widespread amongst medical researchers, leading to inefficient analysis and the potential for important predictive relationships to be missed. Although there can be many competing reasons for dropout, the available joint modelling methods typically allow for only one failure type for the dropout process. We discuss extending the joint modelling methodology to allow for dropout process in a competing risk setting. The standard analyses and joint modelling methodologies will be reviewed through application of data from the HTA funded MAGNETIC trial (a randomised, placebo controlled trial of nebulised Magnesium Sulphate in acute severe asthma in children) to investigate the longitudinal outcome of Yung Asthma Severity Score (ASS) which was measured at baseline and 20, 40, 60, 120, 180 and 240 minutes following randomisation. The reasons for dropout were sometimes clearly related with the good or poor status of the child, or unlikely to be related to the health status but in many instances these reasons were not known or unclear.

Medical Research Council Conference on Biostatistics
in celebration of the MRC Biostatistics Unit's Centenary Year

24th - 26th March 2014 | Queens' College Cambridge, UK

Comparison of Random Forest and parametric imputation models when imputing missing data using MICE: a CALIBER study

Authors:
Shah, Anoop Dinesh (presenting)
Farr Institute of Health Informatics Research at UCL Partners, London
a.shah@ucl.ac.uk

Bartlett, Jonathan
Department of Medical Statistics, London School of Hygiene and Tropical Medicine

Hemingway, Harry
Farr Institute of Health Informatics Research at UCL Partners, London

Poster Session **17.15-19.00** **Old Hall and Kitchens**

Keywords:
Missing data
Missing data
Regression tree
Simulation
Survival

Abstract:
Background: Multivariate Imputation by Chained Equations (MICE) is commonly used for imputing missing values of predictor variables. The 'true' imputation model may contain non-linearities which are not included in default imputation models. Random Forest is an algorithm which combines the predictions of multiple regression trees and should be able to accommodate non-linearities and interactions automatically.
Aim: To develop a Random Forest-based MICE algorithm and compare its performance to parametric MICE.
Methods: We developed methods for multiple imputation using Random Forest within the MICE framework. For continuous variables, we took random draws from independent normal distributions with means predicted using Random Forest. For categorical variables, we imputed missing values from single classification trees based on bootstrap samples of the observed data. We tested these methods in 1000 random samples of 2000 patients drawn from 10,128 patients with stable angina and completely recorded covariates in the CALIBER database of linked electronic health records. We artificially made some variables 'missing at random', and compared the bias and efficiency of coefficient estimates for a survival model after imputation using parametric or Random Forest MICE. We also tested both methods in simulated data with a non-linear association between the fully observed and partially observed predictor variables, but no non-linear associations between predictor variables and survival.
Results: In the CALIBER angina simulation, both methods produced unbiased estimates of log hazard ratios but Random Forest was more efficient and produced narrower confidence intervals. In the second simulation, the coefficient estimate for the partially observed variable was 10% biased using parametric MICE, 2.6% biased using Random Forest with 100 trees and only 1.0% biased using Random Forest with 10 trees.
Conclusions: Random Forest may be useful for imputing complex epidemiological datasets, and may avoid the need to explicitly specify non-linear associations in imputation models for covariates.

Medical Research Council Conference on Biostatistics
in celebration of the MRC Biostatistics Unit's Centenary Year

24th - 26th March 2014 | Queens' College Cambridge, UK

Taking into Account Strains Heterogeneity in the Estimation of Vaccine Efficacy Against Seasonal Influenza

Authors:
Benoit, Anne (presenting)
Institut de Statistique, Biostatistique et Sciences Actuarielles (ISBA), Université catholique de Louvain, Belgium
anne.benoit@uclouvain.be

Legrand, Catherine
Institut de Statistique, Biostatistique et Sciences Actuarielles (ISBA), Université catholique de Louvain, Belgium

Dewé, Walthère
GSK Biologicals, Belgium

Poster Session **17.15-19.00** **Old Hall and Kitchens**

Keywords:
Seasonal Influenza
Design
Tolerance interval
Strain heterogeneity
Vaccine efficacy

Abstract:
Influenza is an infectious disease caused by several virus strains whose repartition varies between geographical regions and seasons. Typically, a vaccine contains 3 or 4 influenza strains and the antigen content is annually reconsidered based on the WHO recommendation. For the same vaccine formulation, pharmaceutical regulations only require efficacy again clinical disease to be shown for a single season, which is performed through a large phase III trial. Subsequent annual modifications of the strain related portion of the vaccine only have to be validated through immunogenicity trials. Classically, influenza vaccine efficacy (VE) trials take place over a single season but over several geographical regions assuming common VE. However, depending on the circulating strains characteristics such as their immunogenicity and their matching levels with the vaccine strains, the vaccinal protection level may vary from one season/region to another.
We argue that not taking this into account provides incomplete and unreliable response as for the benefit of the vaccine in the future. We therefore propose to run phase III VE trials over several geographical regions and seasons in order to characterize the VE heterogeneity. We consider VE as the sum of a common quantity to all clusters (season and geographical region) and of a random cluster-specific part in an adapted meta-analysis approach. VE is then reported as an interval and inference is performed with a one-sided lower tolerance interval, taking into account the heterogeneity across clusters. Such information provides insight on the range of future VE across seasons and geographical regions, instead of a mean past season specific vaccine effect.
Our work will be illustrated by discussing real data examples and simulation results.

Medical Research Council Conference on Biostatistics
in celebration of the MRC Biostatistics Unit's Centenary Year

24th - 26th March 2014 | Queens' College Cambridge, UK

A comparison of multiple imputation methods for bivariate hierarchical data

Authors:
Diaz-Ordaz, Karla (presenting)
London School of Hygiene and Tropical Medicine
Karla.Diaz-Ordaz@lshtm.ac.uk

Kenward, Michael G.
London School of Hygiene and Tropical Medicine

Grieve, Richard
London School of Hygiene and Tropical Medicine

Poster Session **17.15-19.00** **Old Hall and Kitchens**

Keywords: Missing data Hierarchical data Multiple imputation Bivariate outcomes

Abstract:

Background: Missing data are common in cluster randomised trials (CRTs) and for valid inferences need to be handled appropriately. When multiple imputation (MI) is used, the imputation model must recognise the data structure.
We compare complete case analysis, random–effects MI, fixed-effects MI, and single-level MI, when the analysis model is a linear mixed-model.

Methods: We conducted a simulation study to assess the performance (bias and confidence interval (CI) coverage) of the alternative methods. We simulated cluster randomised trial data, consisting of bivariate continuous outcomes with baseline individual and cluster-level covariates. Missing-at-random data scenarios were simulated following a full-factorial design. Amongst the simulation factors were ICCs, number and size of clusters, proportion of missing data, and whether the missing mechanism was associated with treatment. There were 192 scenarios. An Analysis of Variance was carried out to study the influence of the simulation factors and their interactions (up to 4-way interactions) on each performance measure.

Results: Complete case analyses resulted in biased treatment effect estimates (percentage bias between 22% and 60%) especially when treatment was associated with missingness, while multilevel MI produced estimators with negligible bias across all the missing mechanisms considered (percentage bias range 0,003%, 3.35%) and led to CI coverage levels of approximately 95% throughout. Fixed-effect MI over-estimated the SEs, resulting in CI coverage in excess of nominal levels (up to 100%) and was shown to introduce bias in certain scenarios. The most influential factor for bias and coverage was the MI method, but within each MI method, ICC and the number and size of the clusters was seen to have the highest impact on bias and coverage.

Conclusion: Estimates may differ depending on how the MI accommodated for clustering. Multilevel MI performed well across the settings considered and is appropriate for studies that have a hierarchical design.

Medical Research Council Conference on Biostatistics
in celebration of the MRC Biostatistics Unit's Centenary Year

24th - 26th March 2014 | Queens' College Cambridge, UK

Generalising results from randomised controlled trials using linked electronic healthcare data

Authors:
Harron, Katie (presenting)
University College London, Institute of Child Health
katie.harron.10@ucl.ac.uk

Gilbert, Ruth
University College London, Institute of Child Health

Wade, Angie
University College London, Institute of Child Health

Poster Session **17.15-19.00** **Old Hall and Kitchens**

Keywords: data
linkage
generalizability
routine data
infection
paediatric intensive care

Abstract:
Objectives: The CATCH trial (CATheters in CHildren) will determine the effectiveness of impregnated central venous catheters (CVCs) compared with standard CVCs for preventing blood stream infection (BSI) in paediatric intensive care (PICU). Understanding the generalisability (or external validity) of results to PICUs across the NHS requires information on infection rate trends, taking into account on-going improvements in infection control over recent years. Linked electronic healthcare data could allow estimation of the absolute risk difference (potential BSI avoided) if impregnated CVCs were adopted for all PICUs across the NHS.
Methods: PICU admission data (PICANet) for England and Wales from 2003-2012 were linked with laboratory surveillance data collected by Public Health England. Multi-level Poisson regression was used to identify "trial effects" and to estimate risk-adjusted BSI rates. CVC use was identified using a predictive model based on CVC audit data. The number of BSI avoided if impregnated CVCs were adopted (estimated by applying the relative-risk to the baseline risk at the end of the trial) was compared with the total number of CVCs required across PICU (estimated from survey data).
Results: BSI rates in children with CVCs decreased by an average 9% (95% CI 8-12%) per year from 2003-2012 and were higher in trial participants compared with non-participants (IRR 1.78; 1.29-2.48). In 2012, the baseline risk of BSI in children with CVCs was 4.27 per 1000 bed-days (3.74-4.80). A relative-risk of 0.18 (derived from meta-analysis) would have resulted in 203 fewer BSI using impregnated vs standard CVCs. Assuming 8958 CVCs were used in 2012, impregnated CVCs would be cost-effective if on average, each BSI avoided saved costs of at least £1500.
Conclusions: Linked electronic healthcare data can help determine the generalisability of trial results and inform purchasing decisions and implementation for units most likely to benefit. **Conclusion** Estimates may differ depending on how the MI accommodated for clustering. Multilevel MI performed well across the settings considered and is appropriate for studies that have a hierarchical design.

Medical Research Council Conference on Biostatistics
in celebration of the MRC Biostatistics Unit's Centenary Year

24th - 26th March 2014 | Queens' College Cambridge, UK

Using Propensity Score to adjust for unmeasured confounders in small area studies

Authors:
Wang, Yingbo (presenting)
MRC-PHE Centre for Environment and Health, Imperial College London
y.wang11@imperial.ac.uk

Blangiardo, Marta
MRC-PHE Centre for Environment and Health, Imperial College London

Best, Nicky
MRC-PHE Centre for Environment and Health, Imperial College

Poster Session **17.15-19.00** **Old Hall and Kitchens**

Keywords:
Propensity Score
Missing data
Dirichlet Process
Unmeasured confounders
Small area studies

Abstract:
Small area studies are commonly used in epidemiology to assess the impact of risk factors on health outcomes when data are available at the aggregated level. However the estimates are often biased due to unmeasured confounders which are not taken into account. Integrating individual level information into area level data in ecological studies may help reduce bias. To investigate this, we develop an area level propensity score (PS) to integrate individual-level data and synthesize the ecological PS with the routinely available area level datasets, such as hospital episode statistics. This framework comprises three steps:

1) Individual level survey data is used to obtain information on the potential confounders, which are not measured at the area level. Through a Bayesian hierarchical framework we synthesize these variables and calculate the PS at the ecological level, taking into the account the correlation among the potential confounders.
2) As real survey data are typically characterized by a limited coverage compared to small area studies, we impute the ecological PS in the areas with no survey coverage. We present a simulation study to compare the performance of regression tree and Dirichlet Process in predicting the missing PS when there is possible non-linearity and heterogeneity among the confounders.
3) We include observed PS and imputed PS as a scalar quantity in the regression model linking environmental exposure and health outcome. As the PS has no epidemiological interpretation, we specify the adaptive splines underpinned by RJMCMC to allow for non-linear effects.

We conclude that integrating individual level data via PS is a promising method to reduce the bias intrinsic in ecological studies due to unmeasured confounders and we briefly introduce a real application on small area studies for evaluating the effect of air pollution on CVD/asthma hospital admissions in England.

Medical Research Council Conference on Biostatistics
in celebration of the MRC Biostatistics Unit's Centenary Year

24th - 26th March 2014 | Queens' College Cambridge, UK

Recombination hotspots associated with active chromatin modifications in Arabidopsis for natural genetic polymorphism data

Authors:
Zhao, Xiaohui (presenting)
Department of Plant Sciences, University of Cambridge
xz289@cam.ac.uk

Kelly, Krystyna
Department of Plant Sciences, University of Cambridge

Henderson, Ian
Department of Plant Sciences, University of Cambridge

Poster Session **17.15-19.00** **Old Hall and Kitchens**

Keywords:
Coalescent analysis
Recombination
DNA motifs

Abstract:
Mammalian and yeast recombination hotspots are specified by different genetic and epigenetic mechanisms. To investigate recombination hotspots in plants, we used coalescent analysis of genetic variation to estimate the population scaled recombination rate between pairs of SNPs using two different population data sets (80 Eurasian accessions and 180 Swedish accessions) of *Arabidopsis thaliana*. Using the 80 Eurasian accessions, we found that transcription start sites (TSS) and termination sites are enriched for recombination hotspots and that 'hot promoters' were associated with active chromatin modifications, such as H2A.Z and histone H3 lysine 4 trimethylation (H3K4me3), Low Nucleosome Density and low DNA methylation In contrast to humans, where PDRM9 directs recombination hot spots to an intergenic 7-mer CCTCCCT motif, we found that A-rich and CTT-repeat DNA motifs occurred upstream and downstream of hot TSS. [1] We hypothesise that the specification of hotspots by chromatin is the ancestral state and the use of PDRM9 evolved more recently. Interestingly, we found a strong correlation in the pattern of recombination between the two population data sets, suggesting conservation of recombination hot spots during evolution.

[1] Choi K, Zhao X, Kelly KA et al, Nature Genetics, published online 22 September 2013

Medical Research Council Conference on Biostatistics
in celebration of the MRC Biostatistics Unit's Centenary Year

24th - 26th March 2014 | Queens' College Cambridge, UK

Missing data in musculoskeletal trials: A comparison of methods based on a simulation study

Authors:
Joseph, Royes (presenting)
Research Institute for Primary Care & Health Sciences, Keele University
r.joseph@keele.ac.uk

Ogollah, Reuben
Research Institute for Primary Care & Health Sciences, Keele University

Lewis, Martyn
Research Institute for Primary Care & Health Sciences, Keele University

Poster Session **17.15-19.00** **Old Hall and Kitchens**

Keywords:
missing data
complete case analysis
LOCF
MMRM
Multiple imputation

Abstract:
In a review of 91 published RCTs in arthritis and musculoskeletal conditions in 2010-11, we found that complete case analysis (CCA) and single imputation – such as last observation carried forward (LOCF) – are still the most commonly used approaches to analysis of the primary endpoint in spite of many recommendations towards advanced methods such as mixed models for repeated measures (MMRM) or multiple imputation (MI). The findings indicate a possible belief among researchers that if the dropout rate is low and/or equal between treatment arms, bias is not a concern and advanced methods to handle dropouts are unnecessary. Sample size calculations were often performed without adjustment for expected attrition rate and correlation between baseline and outcome. In this study we performed a detailed simulation aimed to extensively assess and compare the performance of analysis methods under various hypothetical clinical trial scenarios. The use of LOCF led to a biased estimation in all scenarios. CCA produced biased estimates of the treatment effect under missing at random and missing not at random
(MNAR). MMRM and MI provided largely unbiased estimates except in situations where the missingness mechanism was MNAR. The direction and magnitude of the bias was influenced by the direction of dropout (whether the observed outcome observation is good or bad in control or experimental arm) and dropout rate; it did not matter whether or not the dropout rate was equal. In MI, the increase in SE of the estimate due to missing data varied by the direction of dropout. However, the direction of dropout did not influence the SE in MMRM. Sample size calculated based on a method which adjusts for correlation and attrition was effective in retaining the nominal power in MMRM unless the missingness mechanism is MNAR. However, MI failed to retain the nominal power especially when the direction of dropout was different.

Creating a pregnancy register for a UK population

Authors:
Boggon, Rachael (presenting)
Clinical Practice Research Datalink

Williams , Tim
Clinical Practice Research Datalink

Poster Session **17.15-19.00** **Old Hall and Kitchens**

Keywords: Pregnancy
Pharmcoepidemiology
Databases
Linkages
Validation

Abstract:
Increased focus is being placed on in-utero exposures with respect to pregnancy outcomes and conditions manifested in children, including studies into teratogenic effects of drugs, which are challenging to investigate using traditional pre-marketing methodologies for obvious reasons. The Clinical Practice Research Datalink (CPRD) provides GP data (GOLD) linked to hospital and mortality data, within which in-utero exposure and maternal and child outcomes can be identified. The objectives are:

(i) Create a pregnancy register using CPRD GOLD
(ii) Incorporate linked data to maximise effectiveness and utility
(iii) Validate results

Women with delivery records in the CPRD GOLD and linked databases were identified. Babies born after the start of prospective data collection with their own records in CPRD GOLD were identified. A cartesian join of mothers to babies by family number was undertaken and the optimal pair identified to create a mother to live baby link. Women's records in all three databases were searched for evidence of pregnancy to create the full pregnancy register. Validation of the register was conducted via free text searches and GP questionnaires.

The mother to live baby link contains approximately 700,000 mothers and over 1 million babies. However, these records account for under 30% of pregnancies in the register. Identifying additional pregnancies, including mothers whose live births were not linked to baby records and those with other birth outcomes, results in a register that can be used for a broader range of pregnancy related research. Full results of the pregnancy register, incorporation of linked databases and the validation exercises will be presented.

The CPRD pregnancy register allows studies of the teratogenic effects of drugs to be conducted in a UK population. Long follow-up allows for studies that require information in all three trimesters, at delivery, through infancy and into childhood.

Medical Research Council Conference on Biostatistics
in celebration of the MRC Biostatistics Unit's Centenary Year

24th - 26th March 2014 | Queens' College Cambridge, UK

Analysis of serial cross-sectional data: incidence of blood-borne viruses in people who inject drugs

Authors:
Harris, Ross J (presenting)
Public Health England
ross.harris@phe.gov.uk

De Angelis, Daniela
MRC Biostatistics Unit, Cambridge

Farrington, Paddy
Open University

Poster Session **17.15-19.00** **Old Hall and Kitchens**

Keywords:
force of infection
seroprevalence surveys
people who inject drugs
frailty

Abstract:
The analysis of cross-sectional data on age-specific prevalence has a long history, dating back to Muench's "catalytic model" in 1934. Such methods aim to estimate the age-specific rate at which susceptible individuals acquire infection; and where data are available at multiple time points, both age- and time-specific effects may be considered. If data are available on more than one infection (transmitted via a common route), the association between them allows individual heterogeneity to be incorporated in the model.

These methods may be applied to cross-sectional data on prevalence of blood-borne infections in people who inject drugs, with consideration of the specific risk patterns in this population. Of particular interest is whether individual heterogeneity may explain the apparent high risk in recent initiates; and changes in risk over time.

Using serological survey data from 1990 to 2012, we examine incidence of hepatitis C and B (HCV, HBV) infection. Incidence is estimated to be far higher in the first six months of injection, then broadly constant: in 2000-2005, HCV incidence was estimated to be 0.227 in the first six months then 0.035 per year; and HBV 0.285 then 0.017 per year. Estimated incidence in the first six months increased over time but incidence thereafter was broadly constant for HCV. HBV incidence declined, largely due to increases in vaccination, with incidence remaining stable in those not vaccinated.

Analysis of paired data revealed some evidence of individual heterogeneity, with around 10% of PWIDS having a 2-fold increase in risk under a gamma frailty model. Further investigation of the role of HBV vaccination is required and we discuss potential extensions to this approach; and other factors such as time-varying heterogeneity. Implementation within a Bayesian framework may allow more complex models to be formulated, and the incorporation of other evidence within the model.

Change Point Semi-parametric mixed effects models for longitudinal data analysis of large population surveys

Authors:
Boukouvalas, Alexis (presenting)
Non-linearity and Complexity Research Group, Aston University
boukouva@aston.ac.uk

Farah, Marian
MRC Biostatistics Unit, University of Cambridge

Nabney, Ian T.
Non-linearity and Complexity Research Group, Aston University

Poster Session **17.15-19.00** **Old Hall and Kitchens**

Keywords:
Longitudinal
Gaussian Process
Dirichlet Process
Change Point

Abstract:
Large longitudinal datasets are becoming increasingly prevalent in areas such as social sciences and genomics. The Linear Mixed Effects model (LME) is the standard model used in the analysis of such datasets due its ease of use and interpretability. However a large number of simplifying assumptions implicit in the LME such as linearity of effects and normality of the model coefficients are rarely justified in practice. We develop a novel method that relaxes both assumptions whilst maintaining a high degree of interpretability. A Dirichlet Process mixture prior is placed on the model coefficients and a change point model is used to model the effect of covariates on the time series model. Further we assume there is a function f(x) which maps the covariates to the location of the change point. Rather than specifying a parametric function for f(x), a Gaussian Process prior is placed on f(x) allowing the description of complex factor effects on the change point locations. In addition to sampling, we explore variational inference methods that allow model inference to be practicable for large datasets. The resulting model provides interpretable results on the effects of covariates on the response and allows for the incorporation of prior judgments concerning characteristics of the population and factor effects. We apply the model on the English Longitudinal Survey of Ageing (ELSA), a large longitudinal survey of older adults aged over 50 in the United Kingdom. We discuss the different covariate effects that affect quality of life and cognitive function of older individuals in the ELSA dataset.

Medical Research Council Conference on Biostatistics
in celebration of the MRC Biostatistics Unit's Centenary Year

24th - 26th March 2014 | Queens' College Cambridge, UK

Bayesian evidence synthesis and combining randomized and nonrandomized results: a case study in diabetes

Authors:
Verde, Pablo Emilio (presenting)
Coordination Centre for Clinical Trials, University of Dusseldorf, Germany
pabloemilio.verde@uni-duesseldorf.de

Ohmann, Christian
Coordination Centre for Clinical Trials, University of Dusseldorf, Germany

Poster Session **17.15-19.00** **Old Hall and Kitchens**

Keywords:
cross-design synthesis
DAG
MCMC
Bias-modelling

Abstract:
There is an increasing interest in combining results from randomized control trials (RCTs) and non-randomized studies in evidence synthesis. One motivation is the generalization of results of randomized control trials in clinical practice, in particular to a group of patients which may not be included in RCTs for ethical reasons.
In this work we present a Bayesian hierarchical model for combining results from different study types. The model explicitly includes two types of parameters: those which are the focus of inference (e.g. treatment effect) and those which are used to describe the data collection processes. These data collection processes parameters are used to directly model potential sources of bias or inconsistencies between sources of evidence.
We illustrate the statistical approach with a real example in diabetes which includes three sources of evidence: aggregated results, partially observed results and patient individual data from a single study's arm. Three types of bias are including into the model: internal validity bias, ecological bias and selection bias. The focus of inference is to extrapolate the RCT's treatment effect to a subgroup of patients of the cohort study which could not meet RCT's inclusion criteria requirements. Inconsistency between sources of evidence is analyzed and prediction to the subgroup of patients is performed by combining parameters from both sources of evidence.

Medical Research Council Conference on Biostatistics
in celebration of the MRC Biostatistics Unit's Centenary Year

24th - 26th March 2014 | Queens' College Cambridge, UK

Binary matched pairs

Authors:
Kartsonaki, Christiana (presenting)
University of Oxford
christiana.kartsonaki@oncology.ox.ac.uk

Poster Session **17.15-19.00** **Old Hall and Kitchens**

Keywords:
matched pairs
conditional logistic regression
efficiency
variance

Abstract:
Consider individuals which are paired, the pairing usually being such that the two individuals in each pair tend to be similar. In each pair one individual is assigned at random to treatment 0 and the other to treatment 1. On each individual a binary response is observed, such that the possible observations on a pair, that for group 0 being written first, are (0, 0), (0, 1), (1, 0) and (1, 1). For continuous responses, the matching results in a reduction of variance of the estimate of the treatment effect θ. For binary matched pairs, the treatment effect θ is estimated using a conditional analysis which uses only the discordant pairs (Cox, 1958). An alternative is to ignore the matching and estimate the treatment effect using an unconditional analysis which uses all pairs. The issue is partly concerned with the informativeness of the concordant pairs. There is a major difficulty in such comparisons in ensuring the comparability of parameters when comparisons are made between different nonlinear models for the same data. Robinson and Jewell (1991) showed that adjusting for nonconfounding covariates in logistic regression never results in a reduction in the variance of the estimate of the treatment effect. The efficiency of the conditional and unconditional analysis is compared. In some cases the variance of the estimate of the treatment effect of the unconditional analysis is smaller than that in the conditional analysis.

Medical Research Council Conference on Biostatistics
in celebration of the MRC Biostatistics Unit's Centenary Year

24th - 26th March 2014 | Queens' College Cambridge, UK

Error Estimation in Drug Combination Evaluation Using Bayesian Approach

Authors:
Wu, Xikun (presenting)
Amgen Ltd
xikunw@amgen.com

Su, Cheng
Amgen Ltd

Sabin, Tony (presenting)
Amgen Ltd

Poster Session **17.15-19.00** **Old Hall and Kitchens**

Keywords:
drug interaction
combination activity
Loewe Addictive
complex nonlinear equation
Bayesian

Abstract:
One common way to evaluate drug interaction in a combination is to compare the combination activity to a reference activity (f) calculated based on a reference model. Our aim is to use Loewe Addictive (LA) reference model to estimate reference activity f and assess its variability. The calculation of f is through solving a nonlinear equation in which the estimated single drug dose response curves are treated as constant. The challenge is the propagation of error in dose response parameter estimates through solving a complex nonlinear equation. Here we provide a Bayesian based methodology to evaluate the estimation error in term of the standard deviation off.

Heart failure hospitalisations following emergency admission – a comparison of resource-use models

Authors:
Goudie, Rosalind (presenting)
Imperial College London
r.goudie@imperial.ac.uk

Bottle, Alex
Imperial College London

Poster Session **17.15-19.00** **Old Hall and Kitchens**

Keywords:
count data
readmissions
bed-days
heart failure
resource-use

Abstract:
Motivation: Heart failure (HF) is a chronic condition where emergency hospital admission is common. Hospital readmission rates among patients with HF are high, as are repeat readmissions and extended hospital stays. Greater ability to predict frequent readmissions and the number of emergency bed-days post discharge from index admission could potentially identify ways to reduce costly acute and emergency readmissions.
Methods: Using three years of data from Hospital Episode Statistics (HES), we considered different approaches for modelling hospital resource-use in the year following an index emergency admission with a primary diagnosis of heart failure. The performance of models for the number of emergency readmissions, were compared with those for the number of unplanned hospital bed-days, and those where the number of bed-days are first categorised into 'resource-use buckets'. As comorbidities are common among HF patients, emergency readmissions with a primary diagnosis of heart failure and all-cause emergency readmissions were treated separately.
A range of count data models, including zero-inflated and hurdle models, were employed for modelling number of readmissions and bed-days. Logistic and multinomial logistic models were used for predicting membership of a particular resource-use bucket. Robust versions of the above models were also considered to account for clustering of patients within hospitals.
Results: The presence of excess zeros and over-dispersion in the data led to the zero-inflated and hurdle models performing better than Poisson and Negative Binomial models. Some, but not all, predictors of number of emergency readmissions were found to also be predictive for number of bed-days in the 12 months following discharge from index admission. Logistic models for the highest resource-use bucket in terms of readmissions or emergency bed-days performed better than models for predicting first readmission.

Medical Research Council Conference on Biostatistics
in celebration of the MRC Biostatistics Unit's Centenary Year

24th - 26th March 2014 | Queens' College Cambridge, UK

Purposeful penalization in Cox regression analyses with less than ten events per variable

Authors:
Heinze, Georg (presenting)
Medical University of Vienna, Centre for Medical Statistics, Informatics and Intelligent Systems, Vienna, Austria
georg.heinze@meduniwien.ac.at

Poster Session **17.15-19.00** **Old Hall and Kitchens**

Keywords:
Cox regression
Penalized likelihood
Bias reduction
Ridge regression
Firth correction

Abstract:
Often biostatisticians are asked to estimate effects of novel biomarkers in survival studies, while adjusting for a relatively large set of variables, e.g., well-established clinical predictors. Often, such studies suffer from a low number of observed events because the disease under study is rare, or patients have relatively good prognosis. In consequence, the number of events per variable (EPV) is low, which may induce bias away from zero and low precision of regression coefficients obtained by multivariable Cox regression.

We investigate whether and how likelihood penalization techniques may improve on those undesirable properties of maximum likelihood estimates. Generally, penalization may serve different purposes, e.g., variable selection, optimization of predictive accuracy, or reduction of the bias of regression coefficients. Popular penalization methods are L1-norm regularization (LASSO), L2-norm regularization (ridge regression), some variants and combinations thereof such as the elastic net, and Firth-type penalization, i.e., penalization based on the Jeffreys prior.

We review and compare ridge regression and Firth-type penalization in the context of studies with a low EPV ratio, but will always consider EPV>1. Furthermore, we investigate ways to tune and combine these methods for the purpose of obtaining the least biased and most efficient adjusted regression coefficients. We present a simulation study comparing the small-sample properties of the procedures, investigating scenarios where only 15-50 events are expected, while there are 5-15 independent variables. Promising results are obtained by a combination of the Firth and the ridge penalty, if the tuning parameter of the ridge penalty is optimized using a criterion that reflects the purpose of the penalization, which in our case is bias reduction. We also present a comparative analysis of a real cancer study, elucidating aspects of practical application of penalization techniques. By way of conclusion, the investigated likelihood penalization methods can help to reduce bias and increase efficiency of maximum likelihood estimates in analyses of medical studies with a critical EPV ratio. Therefore, likelihood penalization should routinely be considered in such cases.

Medical Research Council Conference on Biostatistics
in celebration of the MRC Biostatistics Unit's Centenary Year

24th - 26th March 2014 | Queens' College Cambridge, UK

A Residual Centring approach for the analysis of maternal antenatal anxiety, postnatal stroking and emotional problems in children

Authors:
Hellier, Jennifer May
King's College London
jennifer.hellier@kcl.ac.uk

Poster Session **17.15-19.00** **Old Hall and Kitchens**

Keywords:
Orthagonalizing
Residual centring
Weighted analyses
prenatal anxiety

Abstract:
Residual centering (i.e. orthogonalizing) is a comparable alternative to the mean centering approach for the modelling of interaction terms, which also serves to eliminate multicollinearity in regression analysis. Residual centering is based on a two stage ordinary least squares (OLS) procedure in which the interaction term is regressed onto its representative first order effects and confounders. The residuals of this regression are then used in replacement of the product term (Little, Bovaird et al. 2006). By centring and residualizing the interaction terms, lower order interaction and main effects remain interpretable as average effects.
We apply this methodology to investigate whether the effects of maternal stroking can modify the association between prenatal depression and physiological and child emotional reactivity at 2.5 years of age. Given animal and human evidence for sex differences in the effects of prenatal stress we also compare associations in boys and girls.
The data for this analysis comes from a general population sample of 1233 first time mothers recruited at 20 weeks gestation, from which a random sample of 316 were taken for assessment at 32 weeks, stratified inter-partner psychological abuse, a risk indicator for child development. Of these mothers, 243 reported how often they stroked their infants, and completed the Child Behavior Checklist (CBCL) at 2.5 years post-delivery.
This poster will present the results of these analyses. We report estimates for the general population from the stratified sub-sample by the use inverse sampling probability weights. Weights took account not only of the original stratification but also of the sample attrition. Variation in the weights associated with the covariates of each model was removed to improve efficiency. We demonstrated a significant interaction between prenatal anxiety and maternal stroking in the prediction of CBCL internalizing symptoms.

Little, T. D., J. A. Bovaird, et al. (2006). "On the merits of orthogonalizing powered and product terms: Implications for modeling interactions among latent variables." Structural Equation Modeling-a Multidisciplinary Journal **13**(4): 497-519.

Hierarchical/sequential analysis of time-course flow cytometry data with Dirichlet process mixture modelling

Authors:
Hejblum, Boris (presenting)
University Bordeaux, INSERM, Centre INSERM Bordeaux, France
Boris.hejblum@isped.u-bordeaux2.fr

Thiébaut, Rodolphe
University Bordeaux, INSERM, Centre INSERM Bordeaux, France

Caron, François
Department of Statistics, University of Oxford

Poster Session **17.15-19.00** **Old Hall and Kitchens**

Keywords:
Nonparametric Bayesian
Longitudinal Data
Dirichlet process mixture models
Flow-cytometry

Abstract:
Flow cytometry is a high-throughput technology used to quantify multiple surface and intracellular markers at the level of a single cell. Improvements of this technology lead today to the ability of describing millions of individual cells from a blood sample, across multiple markers. In clinical trials, such measurements can now be performed across several time points. This results in huge datasets, whose manual analysis is highly time-consuming and poorly reproducible. Furthermore, manual analyses are performed to quantify specific populations of interest, possibly missing out other populations.
Several methods have been developed to perform an automatic recognition of cell populations based on flow cytometry raw data. A large part of published applications are actually based on single sample in one patient using up to 4 colors (dimensions) although repeated measurements with several samples by patient by time points are usually available in clinical trials.
We propose to use a Bayesian nonparametric approach with Dirichlet process mixtures (DPM) to model such data. DPMs enable the number of cell populations to be estimated from the data, without resorting to model selection. We propose hierarchical and sequential extensions of Dirichlet process mixtures in order to take into account variability among subjects and over time, and to allow borrowing of information.

Biostatistical Contribution to Medical Research

Authors:
Kanik, Emine Arzu (presenting)
Mersin University, Department of Biostatistics, Mersin, Turkey
arzukanik@gmail.com

Poster Session **17.15-19.00** **Old Hall and Kitchens**

Keywords:
ethics in biostatistics
statistical contribution
MedicReS GBP
Good Biostatistical Practice
Publication Pollution

Abstract:
Background: A case study has been applied to search for the Biostatistical contribution as it is as important as the ghost writing and both threaten the reliability of the medial literature.
Material Method: In this study, 500 research articles published in SCI Indexed Medical Journals addressed in Turkey have been examined and correspondent authors of each has been contacted by e mail. MedicReS GBP "PART III: Publication-Rules and Regulations about Biostatistical Contribution to Medical Research" has been used as method and all articles have been reviewed accordingly in details in conjunction with all the details of this guideline.
Results: In 8% of the articles,there exists biostatistician,statistician and publich health associate.
For the rest, correspondent authors have been asked about the statistical contribution by whom or where?
For the rest, 5% has indicated a paid consultancy service Support and 30% free of charge Support from one of the authors in the Research.none has an information about the statistical capability of the statistical contributors. 24% have been back by the replies of "We paid fort he statistical contribution so we dont have to tell this' or 'one of my friends has a contribution and he did not want his name to appear in the article as an author or at the acknowledgment"
Conclusion: According to MedicReS GBP Part III, in the study, statistical method part in all the Research articles have been observed very poor and insufficient and inefficient in Material-Method Sections. In 95% of the articles, in this section, only general information and general test names have been written. Especially in the developing countries; authors, reviewers and editors have to follow MedicReS Good Biostatistical Practice Guideline for planning, analyzing, reporting and rewiewing their researches to be able to have the quality and reliability.

Medical Research Council Conference on Biostatistics
in celebration of the MRC Biostatistics Unit's Centenary Year

24th - 26th March 2014 | Queens' College Cambridge, UK

On sample size determination for class comparison and predictive classification

Authors:
Czanner, Gabriela (presenting)
Department of Biostatistics and Department of Eye and Vision Science, University of Liverpool
czanner@liv.ac.uk

Cheyne, Christopher
Department of Biostatistics, University of Liverpool

Garcia-Finana, Marta
Department of Biostatistics, University of Liverpool

Poster Session **17.15-19.00** **Old Hall and Kitchens**

Keywords:
sample size
effect size
class comparison
classification
linear discriminant

Abstract:
A class comparison and classification are often used in development of prognostic markers or screening tests. First a class comparison is used to identify which measurements differentiate between the groups of subjects. The classification methods are used to develop a prognostic marker or classifier. The sample size approaches for these two problems should however be different; indeed the second problem does require more data. However, the existing sample size methods for classification are typically based on one measurement taken on a patient i.e. they are univariate; and there is increasing critique for the use of the over-optimistic "rule-of-thumb" formulas. We first summarize the problem of sample size calculation when several measurements are taken on patient. Specifically, we show how the sample size requirements depend on correlation between the predictor variables and the number of predictors. Furthermore we propose simple formula for sample size calculations based on bias-corrected effect size, we study its properties, and show how it can be employed using the data from pilot comparative studies or from published relevant papers. We illustrate the ideas on examples from clinical ophthalmic studies.

Medical Research Council Conference on Biostatistics
in celebration of the MRC Biostatistics Unit's Centenary Year

24th - 26th March 2014 | Queens' College Cambridge, UK

Rapid exploration of parameter-covariate relationships in complex hierarchical models: Application to insulin kinetics in pregnant women with type 1 diabetes

Authors:
Goudie, Robert (presenting)
MRC Biostatistics Unit, Cambridge
robert.goudie@mrc-bsu.cam.ac.uk

Lunn, Dave
MRC Biostatistics Unit, Cambridge

Hovorka, Roman
Department of Paediatrics, University of Cambridge

Poster Session **17.15-19.00** **Old Hall and Kitchens**

Keywords:
Bayesian variable selection
Hierarchical models
Type 1 diabetes
MCMC
BUGS

Abstract:
The parameters of complex individual-level models within hierarchical models may depend on some of the observed covariates. For example, the rate of insulin uptake for individuals might differ across age groups. We aim to explore rapidly the wide range of potential parameter-covariate relationships, aiming to selecting the best predictive model.
Pharmacokinetic models are typically non-linear (in their parameters), often complex, and sometimes only available as a set of differential equations, with no closed-form solution. Fitting such a model for just a single individual can be challenging. Fitting a joint model for all individuals can be even harder, even without the complication of an overarching variable selection objective. We describe a two-stage approach that decouples the variable selection model from each complex individual-level model for the parameters, but nevertheless accounts fully for uncertainty.
In the first stage, we approximate the posterior distribution of each individual's parameters in each individual-level model using Markov chain Monte Carlo (MCMC). In the second stage, we use the samples generated in stage one to form 'proposal distributions' for the individual-level parameters in the full hierarchical model. Using these particular proposal distributions makes stage two very efficient, allowing for rapid exploration of covariate models using reversible jump MCMC. Our approach is implemented within the BUGS software.
We use this methodology to study the factors influencing 4 parameters relating to insulin kinetics in pregnant women with type 1 diabetes (T1D), using data from two clinical studies. We believe that such an analysis would be impracticable without the two-stage methodology. We find a number of factors that influence insulin kinetics in pregnant women with T1D, including gestational age. Our results increase understanding of factors affecting insulin kinetics and contribute to the development of novel personalised approaches such as the artificial pancreas to aid management of pregnant women with T1D.

Medical Research Council Conference on Biostatistics
in celebration of the MRC Biostatistics Unit's Centenary Year

24th - 26th March 2014 | Queens' College Cambridge, UK

A comparison of Bayesian adaptive randomization and multi-stage designs for multi-arm clinical trials

Authors:
Wason, James (presenting)
MRC Biostatistics Unit, Cambridge
james.wason@mrc-bsu.cam.ac.uk

Trippa, Lorenzo
Harvard School of Public Health

Poster Session **17.15-19.00** **Old Hall and Kitchens**

Keywords:
Adaptive randomization
group-sequential designs
multi-arm trials
multiple testing

Abstract:
When several experimental treatments are available for testing, multi-arm trials provide gains in efficiency over separate trials. Including interim analyses allows the investigator to effectively use the data gathered during the trial. Bayesian adaptive randomization (AR) and multi-arm multi-stage (MAMS) designs are two distinct methods that use patient outcomes to improve the efficiency and ethics of the trial. AR allocates a greater proportion of future patients to treatments that have performed well; MAMS designs use pre-specified stopping boundaries to determine whether experimental treatments should be dropped. There is little consensus on which method is more suitable for clinical trials. In this presentation we compare the two designs under several simulation scenarios and in the context of a real multi-arm phase II breast cancer trial. We compare the methods in terms of their efficiency and ethical properties. The practical problem of a delay between recruitment of patients and assessment of their treatment response is also considered. Both methods are more efficient and ethical than a multi-arm trial without interim analyses. Delay between recruitment and response assessment attenuates this efficiency gain. Our comparisons show that AR is more efficient than MAMS designs when there is an effective experimental treatment; while if none of the experimental treatments is effective, then MAMS designs slightly outperform AR.

Medical Research Council Conference on Biostatistics
in celebration of the MRC Biostatistics Unit's Centenary Year

24th - 26th March 2014 | Queens' College Cambridge, UK

Bayesian generalised additive models for estimating influenza-associated mortality rates

Authors:
Presanis, Anne (presenting)
MRC Biostatistics Unit, Cambridge
anne.presanis@mrc-bsu.cam.ac.uk

Goldstein, Ed
Harvard School of Public Health

De Angelis, Daniela
MRC Biostatistics Unit, Cambridge

Poster Session **17.15-19.00** **Old Hall and Kitchens**

Keywords:
Evidence synthesis
Bayesian
Time series
Influenza mortality
Disease burden

Abstract:
The 2009 A/H1N1 influenza pandemic highlighted gaps in the ability of countries to provide timely assessments of influenza burden, limiting both health-care planning during seasonal influenza and preparedness for future pandemics. Relevant data are often fragmented and biased, so burden estimation requires methods able to combine heterogeneous and limited evidence coherently. Different aspects of influenza burden, i.e. type-specific health-care consultation rates, case-severity risks and excess mortality due to influenza, have so far been separately estimated, using respectively: (i) a joint regression model of consultation and virological positivity data; (ii) a synthesis of evidence at different severity levels; (iii) a regression of deaths on type-specific health-care consultation rates. The three models have data and parameters in common so estimates from each should, in theory, be consistent with each other. Synthesising the data in a single combined model should ensure this consistency, allowing information in each to propagate coherently from all the data to the parameters. The synthesis framework allows for detection and resolution of any inconsistency, leading to more accurate assessments of uncertainty in burden estimates.
A first step to this composite model is to develop Bayesian versions of models of type (iii), apportioning mortality rates to a "baseline" rate and excess mortality associated with different influenza types. We develop age-specific Bayesian Normal, Poisson and Negative Binomial regression models with identity link for observed deaths in both the US and the UK. Examination of Bayesian residuals from models not accounting for auto-correlation suggests a latent AR(1) process in the mean is appropriate. Preliminary results suggest models accounting for over-dispersion, whether through auto-regression or through a dispersion parameter, are preferred by the Deviance Information Criterion. However, estimates of the contribution of influenza to mortality are sensitive to the inclusion or not of a latent auto-regressive process. Next steps are to embed these mortality models in larger evidence syntheses.

Medical Research Council Conference on Biostatistics
in celebration of the MRC Biostatistics Unit's Centenary Year

24th - 26th March 2014 | Queens' College Cambridge, UK

Correction of unexpected distributions of P values from analysis of whole genome arrays

Authors:
Barton, Sheila (presenting)
MRC Lifecourse Epidemiology Unit, University of Southampton
S.J.Barton@soton.ac.uk

Crozier, Sarah
MRC Lifecourse Epidemiology Unit, University of Southampton

Inskip, Hazel
MRC Lifecourse Epidemiology Unit, University of Southampton

Poster Session **17.15-19.00** **Old Hall and Kitchens**

Keywords:
Genomics
Statistical assumptions
Distributions of P values
Epigenome

Abstract:
Background In the last 10 years microarrays have become a fundamental tool in biological research laboratories throughout the world to study genotype, gene expression and more recently epigenetic measurements. Statistical analysis of genome-wide microarrays can result in many thousands of identical statistical tests being performed as each probe is tested for an association with a phenotype of interest. If there were no association between any of the probes and the phenotype, the distribution of P values obtained from statistical tests would resemble a uniform distribution. If a selection of probes were significantly associated with the phenotype we would expect to observe P values for these probes of less than the designated significance level, alpha, resulting in more P values of less than alpha than expected by chance.

Results Using data from a whole genome methylation promoter array we tested the association between probes and phenotype using linear regression. We observed P value distributions where there were fewer P values less than alpha than would be expected by chance. Our data suggest that a possible reason for this is a violation of the statistical assumptions required for linear regression arising from heteroskedasticity. A simple but statistically sound remedy rectified this violation and resulted in meaningful P value distributions. A heteroskedasticity–consistent covariance matrix estimator was used to calculate standard errors of regression coefficients that are robust to heteroskedasticity. These methods are readily available in most statistical software packages including Stata, R and SPSS.

Conclusions The statistical analysis of 'omics data requires careful handling, especially in the choice of statistical test. To obtain meaningful results it is essential that the assumptions behind these tests are carefully examined and any violations rectified where possible, or a more appropriate statistical test chosen.

Medical Research Council Conference on Biostatistics
in celebration of the MRC Biostatistics Unit's Centenary Year

24th - 26th March 2014 | Queens' College Cambridge, UK

Repeated observations in longitudinal studies – linear and non-linear mixed-effects models and survival analysis

Authors:
Fritz, Josef (presenting)
Innsbruck Medical University, Department for Medical Statistics, Informatics and Health Economics
josef.fritz@i-med.ac.at

Ulmer, Hanno
Innsbruck Medical University, Department for Medical Statistics, Informatics and Health Economics

Poster Session **17.15-19.00** **Old Hall and Kitchens**

Keywords:
Longitudinal data
Mixed models
Survival analysis

Abstract:

In contrast to clinical trials, longitudinal data generated from observational studies show a higher degree of complexity. Characteristics like unequal numbers of repeated measurements, measurements taken at arbitrary time-points, correlations between the measurements or group- and subject-specific evolutions are occurring frequently. Based on data of the Vorarlberg Health Monitoring & Promotion Programme, we investigated rates of changes in total cholesterol and their influence on survival with a combination of linear and non-linear mixed effects and Cox proportional hazards regression models. Primary aim was the detection of sex and age dependent patterns in total cholesterol and their relation to survival. Intra-individual courses of total cholesterol for men and women at different ages were estimated with fixed and random effects in the longitudinal models and the estimated parameters were subsequently used as an input into the survival models. In doing so, the Cox regression model can be extended in that way that not only the impact of a predictor variable at a single time-point, but also the trend of this predictor over time is accounted for in the analysis of survival.

Medical Research Council Conference on Biostatistics
in celebration of the MRC Biostatistics Unit's Centenary Year

24th - 26th March 2014 | Queens' College Cambridge, UK

Impact of rigorously examining the effectiveness of Naloxone (take-home; on release) to reduce opiate overdose deaths

Authors:
Bird, Sheila (presenting)
MRC Biostatistics Unit
Sheila.Bird@mrc-bsu.cam.ac.uk

on behalf of N-ALI VE Trial's co-principal investigators

Poster Session **17.15-19.00** **Old Hall and Kitchens**

Abstract:
Naloxone, the opiate antidote, can be injected intramuscularly and is used by doctors in accident and emergency or by ambulance staff to reverse opiate overdose.

The very high risk of drugs-related death (DRD) in the fortnight following release from prison – 7 times higher than at comparable other times at liberty; and 1 DRD per 200 released adult male prisoners with a history of heroin injection – was first quantified in Scotland. Bird and Hutchinson, in 2003, proposed a randomized controlled trial of Naloxone-on-release to counter this very high DRD-rate.

In 2008, the Medical Research Council approved funding for the pilot N-ALIVE Trial to randomize, equally in two prison jurisdictions, the first 10% of 56,000 prisoners (with a history of heroin injection) needed to determine if those randomized to receive Naloxone-on-release experience 30% fewer DRDs in the first 4 weeks post-liberation than those randomized to the control group.

As of 2011, however, the N-ALIVE Trial could not randomize in Scotland (where prisons have been specifically resourced to prescribe Naloxone for at-risk prisoners) nor in Wales as take-home Naloxone had become a public health policy in both countries.

I outline how monitoring of Scotland's National Naloxone Programme was designed and subsequently modified; summarize recruitment to the pilot N-ALIVE Trial; give an updated international perspective on take-home Naloxone; and anticipate England's decision on a randomized step-wedge design for its evaluation of take-home Naloxone.

Is it "impact" for a randomized controlled trial to be overtaken by the public policy it seeks to inform?