The International Biometric Society
Channel Network Conference 2015

CONFERENCE PROGRAMME & ABSTRACTS

Nijmegen, 20-22/04/2015
Scientific committee IBS Channel Network Conference 2015:
Aurelien Latouche (Chair; CNAM, Paris, France)
Cécile Proust-Lima (University of Bordeaux, Bordeaux, France)
David Causeur (Agrocampus Ouest, Rennes, France)
Catherine Legrand (UC Louvain, Louvain, Belgium)
Stijn Vansteelandt (University of Ghent, Ghent, Belgium)
Daniel Jackson (MRC Biostatistics, Cambridge, UK)
Diana Cole (University of Kent, Kent, UK)
Jeanine Houwing (Leiden University medical center, Leiden, the Netherlands)
Mark van de Wiel (VU University medical center, Amsterdam, the Netherlands)

Local organizing committee:
Ronald Geskus (Chair; Academic Medical Center, Amsterdam, the Netherlands)
Jelle Goeman (Treasurer; Radboud university medical center, Nijmegen, the Netherlands)
Joanna in ’t Hout (Secretary; Radboud university medical center, Nijmegen, the Netherlands)
Mark van de Wiel (PR; VU University medical center, Amsterdam, the Netherlands)

Sponsors:
Koninklijke Nederlandse Akademie van Wetenschappen (KNAW)
STAR - Stochastics, Theoretical, and Applied Research
Janssen pharmaceutical company of Johnson-Johnson
Danone Nutricia Research
BMS-ANed, Dutch region of the IBS

Typeset in \LaTeX with Springer Verlag’s svmult.cls document class.
Welcome to the 2015 Channel Network Conference.

We are pleased to welcome you to the 5th Channel Network Conference, the biennial scientific meeting of the regions Belgium, France, Great-Britain/Ireland and the Netherlands of the International Biometric Society. This year it will be hosted by the Biometric Section of the Dutch Society for Statistics and Operations Research (http://www.vvs-or.nl) and the Dutch region of the IBS (http://www.bms-aned.nl). A special welcome goes to the members of the IBS executive board who have their board meeting right after our conference.

The Channel Network Conference aims to enhance scientific communication and collaboration at a more local level. We think we have created an interesting program with two short courses, one keynote speaker, three invited sessions, 15 contributed sessions and two sessions with poster lightning presentations. In this abstract book you will find all information on the scientific program.

Please also have a look at the interesting social program, including a poster session with drinks and food on Monday and a conference dinner along the river Waal on Tuesday. Further information on the conference dinner can be found in the conference bag.

A special thanks goes to our sponsors:

- the Royal Netherlands Academy of Arts and Sciences (KNAW; http://knaw.nl)
- the Dutch cluster ”Stochastics -Theoretical and Applied Research” (STAR; http://www.eurandom.tue.nl/STAR)
- Danone Nutricia Research (http://www.nutricia.com)
- Janssen Pharmaceuticals(http://www.janssenpharmaceuticalsinc.com).

We would also like to thank Dennis te Beest, who was of great help in preparing this abstract book.

We hope that you will enjoy the conference, the city of Nijmegen and its surroundings. On behalf of the scientific and local organising committee,

Ronald Geskus
Academic Medical Center (AMC), Amsterdam
A word of welcome
Contents

KEYNOTE

Analysis of complex phenotypes in genomics ............................. 25
Sylvia Richardson

INVITED SESSION: DYNAMIC PREDICTION

Pseudo-observations, landmarking and dynamic prediction ............... 29
Hein Putter

Personalized Screening Intervals for Biomarkers using Joint Models for
Longitudinal and Survival Data .............................................. 30
Dimitris Rizopoulos

On methods for evaluating dynamic predictions ........................... 31
Paul Blanche

INVITED SESSION 2: INTEGRATED POPULATION MODELLING

Building robust abundance indices combining commercial data and scientific
survey ................................................................. 35
Marie-Pierre Etienne, Robyn Forrest, Jean-Baptiste Lecomte and Eric Parent

Differential contribution of vital rate synchrony on population synchrony in
barn swallows ......................................................... 36
Michael Schaub, Johann von Hirschheydt and Martin U. Gruebler

Recent developments in Integrated Population Modelling ................ 37
Takis Besbeas

INVITED SESSION 3: INTEGRATED POPULATION MODELLING

Robust causal discovery in the presence of latent variables ............... 41
Tom Heskes and Tom Claassen

Model misspecification and bias for inverse probability weighting and doubly
robust estimators ......................................................... 42
Ingeborg Waernbaum
Bias-Reduced Doubly Robust Estimation ........................................ 43
Karel Vermeulen and Stijn Vansteelandt

CONTRIBUTED SESSION C1: GENETICS AND OMICS

Testing for splice-QTL effects: finding then SNPs and exons in the haystack . . . 47
Renee de Menezes and Marianne Jonker

Statistical integration of omics data measured by different technology platforms ................................................................. 48
Said el Bouhaddani, Hae-Won Uh, Jeanine Houwing-Duistermaat and Geurt Jongbloed

Reconstruction of genome ancestry blocks in multiparental populations . . . . 49
Chaozhi Zheng, Martin P. Boer and Fred A. van Eeuwijk

GLMMs for nucleic acid concentration estimation in digital droplet PCR . . . . 50
Matthijs Vynck, Olivier Thas and Jo Vandesompele

Measuring telomere lengths using DNA sequencing data: Discovering even more layers of heterogeneity in Prostate Cancer ............................................. 51
Andy Lynch, J Henry Farmery, Charles E Massie and Mike L Smith

CONTRIBUTED SESSION C2: GENETICS AND OMICS

Directed cyclic mixed graph modeling for network integration of high-dimensional genomic data .................................................. 55
Carel F.W. Peeters, Wessel N. van Wieringen and Mark A. van de Wiel

tigaR: Integrative significance analysis of temporal differential gene expression induced by genomic abnormalities ........................................... 56
Viktorian Miok, Saskia M Wilting, Mark A van de Wiel, Annelieke Jaspers, Paula I van Noort, Ruud H Brakenhoff, Peter JF Snijders, Renske DM Steenbergen and Wessel N van Wieringen

Statistical methods to analyze repeated measurements of overdispersed categorical data: an application in longitudinal microbiome data .............. 57
Ivonne Martin, Roula Tsonaka and Jeanine Houwing-Duistermaat

CONTRIBUTED SESSION C3: STATISTICAL MODELLING

Estimating prevalences using medication data and random forests ............... 61
L. Slobbe, K. Fussenich, A. Wong, C. van den Brink, H. Boshuizen, J.J. Polder and H. van Oers
Contents

**Analysis for single patient studies** .......................................................... 62
Paul Garthwaite, Fadlalla Elfadaly and John Crawford

**Bayesian modelling of ordinal agreement** ............................................. 63
Sophie Vanbelle, Alberto Alvarez-Iglesias, John Newell and Carl Scarrott

**CONTRIBUTED SESSION C4: SURVIVAL, PREDICTION, JOINT MODELS**

**Summarising Censored Survival Data Using The Mean Residual Life Function** 67
John Hinde, Alberto Alvarez-Iglesias, John Newell and Carl Scarrott

**Projections of life expectancy and other health indicators and impact of preventive policy for chronic diseases** ................................................. 68
Pierre Joly, Mathilde Wanneveich and Helene Jacqmin-Gadda

**A probabilistic record linkage model for survival data** ......................... 69
Michel Hof and Koos Zwinderman

**CONTRIBUTED SESSION C5: BAYESIAN STATISTICS**

**Bayesian P-spline smoothing using INLA** .............................................. 73
Cajo J.F. ter Braak, María Xosé Rodríguez-Ivarez, Martin Boer, Paul Eilers and Havard Rue

**A dynamic model for the incidence rate of ebola in Sierra Leone** ............ 74
Philippe Lambert

**Bayesian modelling for evidence synthesis via Bayesian melding** ............. 75
Robert Goudie, Anne Presanis, Lorenz Wernisch, Daniela De Angelis and David Lunn

**Using Bayesian Networks (BNs) for predicting patient-reported outcome measures** ................................................................. 76
Maria Gheorghe, Anca Hanea, Melinde R.S. Boland, Maureen Rutten-van Mölken and Pieter H.M. van Baal

**Comparing diagnostic test in meta-analysis: a hierarchical meta-regression approach** ................................................................. 77
Pablo Verde

**CONTRIBUTED SESSION C6: GENETICS AND OMICS**

**Analyzing sets of SNPs via SKAT versus global test: methodical overlaps and their practical consequences** .......................................... 81
Monika Jelizarow and Jelle J. Goeman
**Fast Parametric Warping of Peaks in Metabolomics Data** ........................................... 82
Ron Wehrens, Tom G. Bloemberg and Paul H.C. Eilers

**Augmented prediction in Omics applications** .......................................................... 83
Mar Rodriguez-Girondo and Bart Mertens

**CONTRIBUTED SESSION C7: STATISTICAL MODELLING**

**A demonstration of overfitting and model selection in a unique setting of two independent data sets.** .......................................................... 87
Chamberlain Mbah, Olivier Thas, Jan De Neve, Hubert Thierens, Jenny Chang-Claude, Petra Seibold, Irmgard Helmbold, Sabine Behrens, Odilia Popanda, Peter Schmezer and Kim De Ruyck

**Fast P-splines for very long data series** ................................................................. 88
Martin Boer and Paul Eilers

**Mediation analysis of randomised experiments** ...................................................... 89
Sjouke Vandenberghe and Stijn Vansteelandt

**CONTRIBUTED SESSION C8: SURVIVAL, PREDICTION, JOINT MODELS**

**Bayesian variable selection with joint modeling of a longitudinal and survival outcome assuming different association structures** ........................................... 93
Eleni-Rosalina Andrinopoulou

**Joint modelling of longitudinal and multi-state processes: an application in patients with localized prostate cancer** ......................................................... 94
L. Ferrer, James Dignam, Virginie Rondeau, Tom Pickles and Cecile Proust-Lima

**Joint latent class model for longitudinal data and interval-censored semi-competing events: Application to Alzheimer’s disease** ................................. 95
Anais Rouanet, Pierre Joly, Jean-Francois Dartigues, Cecile Proust-Lima and Helene Jacqmin-Gadda

**CONTRIBUTED SESSION C9: SURVIVAL, PREDICTION, JOINT MODELS**

**Estimating the type-specific cumulative incidence of cervical precancer in a setting with competing risks and masked event types** ........................................... 99
Birgit Witte and Johannes Berkhof

**A unified effect measure for semiparametric linear transformation models** .... 100
Jan De Neve, Thomas A. Gerds and Olivier Thas
Multivariate mixtures of Erlangs for density estimation under censoring and truncation ................................................................. 101
Roel Verbelen

Event-based Ascertainment of Recurrent Events Data ....................... 102
Theodor Adrian Balan, M.A. Jonker, and H. Putter

Dynamic frailty models based on compound birth-death processes ........ 103
Hein Putter and Hans van Houwelingen

Direct modeling of regression effects for transition probabilities in the non-Markov illness-death model ........................................ 104
Leyla Azarang, Thomas Scheike, and Jacobo de Uña-Alvarez-Alvarez

CONTRIBUTED SESSION C10: ECOLOGY

Abundance estimation using integrated modelling of unmarked animal data from camera traps ...................................................... 107
Natoya Jourdain, Diana J. Cole, Martin S. Ridout, and Marcus J. Rowcliffe

Estimation of \( P(X > Y) \) for normal distributions in the context of probabilistic environmental risk assessment for nanoparticles .................. 108
Rianne Jacobs, Andriette Bekker, and Cajo ter Braak

Modelling the sea lice population at a salmon farm .......................... 109
Magne Aldrin

Parameter Redundancy in Integrated Population Models .................... 110
Diana Cole and Rachel McCrea

CONTRIBUTED SESSION C11: SURVIVAL, PREDICTION, JOINT MODELS

Dynamic prediction by direct binomial regression .............................. 113
Mia Klinten Grand and Hein Putter

Dynamic prediction of mortality amongst patients in intensive care using the sequential organ failure assessment (SOFA) score .................. 114
Jammbe Z Musoro

Time dependent discriminative ability of dynamic prediction models with discrete survival times ................................................. 115
N. van Geloven and A.H. Zwinderman

CONTRIBUTED SESSION C12: LONGITUDINAL DATA & CAUSALITY
<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple imputation for incomplete predictors and complex outcomes: mice</td>
<td>119</td>
</tr>
<tr>
<td>vs. sequential Bayesian imputation</td>
<td></td>
</tr>
<tr>
<td>Nicole Erler, Dimitris Rizopoulos, Oscar H. Franco, and Emmanuel M.E.H. Lesaffre</td>
<td></td>
</tr>
<tr>
<td>Pseudo-Likelihood Approach for Large and Complex Ordinal Data</td>
<td>120</td>
</tr>
<tr>
<td>Anna Ivanova, Geert Molenberghs, and Geert Verbeke</td>
<td></td>
</tr>
<tr>
<td>Prediction of glaucoma progression using a high-dimensional two-stage Bayesian hierarchical model</td>
<td>121</td>
</tr>
<tr>
<td>CONTRIBUTED SESSION C13: ECOLOGY</td>
<td></td>
</tr>
<tr>
<td>P-spline models for agricultural field trials</td>
<td>125</td>
</tr>
<tr>
<td>Fred van Eeuwijk, María Xosé Rodríguez-Ivarez, Martin Boer and Paul Eilers</td>
<td></td>
</tr>
<tr>
<td>Latent Block Model for ecological abundance data</td>
<td>126</td>
</tr>
<tr>
<td>Julie Aubert, S. Schbath and S. Robin</td>
<td></td>
</tr>
<tr>
<td>Ultrasonic vocalizing behaviour in socially housed rats: a composite-link function approach</td>
<td>127</td>
</tr>
<tr>
<td>Nadia Vendrig, Ilona Pinter, Suzanne Peters, Berry Spruijt and Cajo ter Braak</td>
<td></td>
</tr>
<tr>
<td>Capture-recapture and the penalized composite link model</td>
<td>128</td>
</tr>
<tr>
<td>Paul Eilers</td>
<td></td>
</tr>
<tr>
<td>CONTRIBUTED SESSION C14: LONGITUDINAL DATA &amp; CAUSALITY</td>
<td></td>
</tr>
<tr>
<td>Using pleiotropic genetic variants to simultaneously estimate multiple causal effects: a multivariable extension to Mendelian randomization</td>
<td>131</td>
</tr>
<tr>
<td>Simon G. Thompson and Stephen Burgess</td>
<td></td>
</tr>
<tr>
<td>Estimating non-linear causal effects from observational data</td>
<td>132</td>
</tr>
<tr>
<td>Ernst C Wit and Mahdi Mahmoudi</td>
<td></td>
</tr>
<tr>
<td>The dynamic approach to causal reasoning in ageing studies</td>
<td>133</td>
</tr>
<tr>
<td>Daniel Commenges</td>
<td></td>
</tr>
<tr>
<td>CONTRIBUTED SESSION C15: STATISTICAL MODELLING</td>
<td></td>
</tr>
<tr>
<td>Fitting linear mixed models with and without centering the explanatory variables</td>
<td>137</td>
</tr>
<tr>
<td>Alan Welsh and Hwan-Jin Yoon</td>
<td></td>
</tr>
</tbody>
</table>
**Contents**

<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable selection and deselection in regression models, with confidence</td>
<td>138</td>
</tr>
<tr>
<td>Aldo Solari and Jelle J. Goeman</td>
<td></td>
</tr>
<tr>
<td><strong>Exact testing with random permutations</strong></td>
<td>139</td>
</tr>
<tr>
<td>Jesse Hemerik and Jelle J. Goeman</td>
<td></td>
</tr>
<tr>
<td><strong>Count data analysis in nutrition clinical trials</strong></td>
<td>140</td>
</tr>
<tr>
<td>Yalcin Yavuz and Sophie Swinkels</td>
<td></td>
</tr>
<tr>
<td><strong>POSTER LIGHTNING PRESENTATIONS P1</strong></td>
<td></td>
</tr>
<tr>
<td>Best-after-breast design: challenges of nutrition intervention studies in infants</td>
<td>143</td>
</tr>
<tr>
<td>Elleny Balder and Sophie Swinkels</td>
<td></td>
</tr>
<tr>
<td>Spline based self-controlled case series method</td>
<td>144</td>
</tr>
<tr>
<td>Heather Whitaker, Yonas Weldesselassie and Paddy Farrington</td>
<td></td>
</tr>
<tr>
<td>Modelling the effect of vaccinating boys along with girls against hpv</td>
<td>145</td>
</tr>
<tr>
<td>Johannes Berkhof and Johannes Bogaards</td>
<td></td>
</tr>
<tr>
<td>Individual patient data meta-analysis for diagnostic test accuracy studies: a review of methods used in practice</td>
<td>146</td>
</tr>
<tr>
<td>Junfeng Wang</td>
<td></td>
</tr>
<tr>
<td>Bayesian Geoadditive Regression and Latent Variable Models of Childhood Malnutrition in Gamo Gofa Zone</td>
<td>147</td>
</tr>
<tr>
<td>Tilahun Ferede Asena and Derbachew Asfaw</td>
<td></td>
</tr>
<tr>
<td>Sensitivity analyses for dependent censoring in survival analysis</td>
<td>148</td>
</tr>
<tr>
<td>Alan Kimber, Stefanie Biedermann and Natalie Staplin</td>
<td></td>
</tr>
<tr>
<td>Survival Analysis Models for MESS Epilepsy Data</td>
<td>149</td>
</tr>
<tr>
<td>Boryana Cristina Lopez Kolkovska</td>
<td></td>
</tr>
<tr>
<td>Calculating a population attributable fraction in the case of clustered data</td>
<td>150</td>
</tr>
<tr>
<td>Hendrik Boshuizen and Marit de Lange</td>
<td></td>
</tr>
<tr>
<td>Measurement error models for accelerometer data using MCMC, Integrated Nested Laplace Approximation and Maximum Likelihood Estimation methods</td>
<td>151</td>
</tr>
<tr>
<td>George O. Agogo, Hilko van der Voet, Pieter van’t Veer, Laura Trijsburg, Fred A. van Eeuwijk and Hendrik Boshuizen</td>
<td></td>
</tr>
<tr>
<td>Dealing with impreciseness in ranked set sampling with fuzzy sets</td>
<td>152</td>
</tr>
<tr>
<td>Bekir Cetintav and Selma Gurler</td>
<td></td>
</tr>
<tr>
<td><strong>POSTER LIGHTNING PRESENTATIONS P2</strong></td>
<td></td>
</tr>
</tbody>
</table>
Enrichment of GWAS data, combining different sources of genetic data
Erik van Iperen, G.K. Hovingh, F.W. Asselbergs and A.H. Zwinderman

Variational Bayesian SEM for undirected Network recovery using external data
Gino Kpogbezan, Mark van de Wiel, Aad van der Vaart, Wessel N. van Wieringen and Gwenael G.R. Leday

Alternative splice variants in gene expression values in patients with Marfans syndrome
Wouter Ouwerkerk and Aeilko H. Zwinderman

Combination approaches improve predictive performance of diagnostic rules for mass-spectrometry proteomic data
Alexia Kakourou, Bart Mertens and Werner Vach

Statistical Methods for Analysing Drug Sensitivity of Cancer Cells
Elaheh Oftadeh and Jian Zhang

Sample size determination for classification using metabolomics data

Accuracy of imputation and prediction in dairy cattle using whole-genome sequence data
Rianne van Binsbergen, Mario P.L. Calus, Marco C.A.M. Bink, Fred A. van Eeuwijk and Roel F. Veerkamp

Expectile smoothing for big data and visualization of linkage disequilibrium decay
Sabine K. Schnabel, F.A. van Eeuwijk and P.H.C. Eilers

Ridge estimation of inverse covariance matrices from high-dimensional data
Wessel van Wieringen and Carel Peeters

SATELLITE MEETING, INFERRING DYNAMIC GENETIC NETWORKS

Chain graphical models for inferring dynamic networks
Fentaw Abegaz

Sparse vector autoregressive models with an application in marketing
Ines Wilms

A focused selection of graphical models in fMRI connectivity studies
Gerda Claeskens

Controversy in mechanistic modelling of biopathways with Gaussian processes
Dirk Husmeier
Realistic network reconstruction methods for the Timing Metabolism (TiMet) research project .............................................. 171
Marco Gregorzyk

Graphical models for causal inference from multivariate time series ............. 172
Michael Eichler

List of Presenters ................................................................. 173
Programme
Monday 20th April

8.30-9.00  Registration, Course
9.00 - 12.15  Course I: Splines, Paul Eilers (Red Room)
9.00 - 12.15  Course II: INLA, Andrea Riebler (Yellow room)
12.15-13.15  Lunch + Registration
13.15-13.30  Opening (Lindenberg room)
13.30-15.30  Afternoon session I
13.30-14.10  Invited I, Dynamic Prediction (organised by Cecile Proust-Lima) (Lindenberg room)
13.30-14.10  Hein Putter
13.10-14.50  Dimitris Rizopoulos
14.50-15.50  Paul Blanche
13.30-13.50  Contributed C1 (Genetics and Omics) (Yellow room)
13.30-13.50  Renee de Menezes
13.50-14.10  Said el Boudaddani
14.10-14.50  Chaozhi Zheng
14.50-15.10  Matthijs Vynck
14.30-14.50  Paul Blanche
14.50-15.30  Andy Lynch
15.30-16.00  Break
16.00-17.00  Afternoon session II
16.00-16.20  Contributed C2 (Genetics and Omics) (Red room)
16.00-16.20  Carel Peeters
16.20-16.40  Viktorian Miok
16.40-17.00  Ivanne Martin
16.00-16.20  Contributed C3 (Statistical Modelling) (Yellow room)
16.00-16.20  Koen Fussenich
16.20-16.40  Paul Garthwaite
16.40-17.00  Sophie Vanbelle
16.00-16.20  Contributed C4 (Survival, prediction, joint models) (Blue room)
16.00-16.20  John Hinde
16.20-16.40  Mathilde Wanneveich
16.40-17.00  Michel Hof
17.00-17.30  Poster lightning presentations
17.00-17.30  Presentations posters P1 (Red room)
17.00-17.30  Elleny Balder
17.00-17.30  Heather Whitaker
17.00-17.30  Johannes Berkhof
17.00-17.30  Junfeng Wang
17.00-17.30  Tilahun Ferede
17.00-17.30  Alan Kimber
17.00-17.30  Boryana Cristina Lopez Kolkovska
17.00-17.30  Henrike Boshuizen
17.00-17.30  George Agogo
17.00-17.30  Bekir Cetintav
<table>
<thead>
<tr>
<th>Time</th>
<th>Name</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>17.00-17.03</td>
<td>Erik van Iperen</td>
<td>Enrichment of GWAS data, combining different sources of genetic data</td>
</tr>
<tr>
<td>17.03-17.06</td>
<td>Gino Kpogbezan</td>
<td>Variational Bayesian SEM for undirected Network recovery using external data</td>
</tr>
<tr>
<td>17.06-17.09</td>
<td>Wouter Ouwerkerk</td>
<td>Alternative splice variants in gene expression values in patients with Marfans syndrome</td>
</tr>
<tr>
<td>17.09-17.12</td>
<td>Alexia Kakourou</td>
<td>Combination approaches improve predictive performance of diagnostic rules for mass-spectrometry proteomic data</td>
</tr>
<tr>
<td>17.12-17.15</td>
<td>Elaheh Oftadeh</td>
<td>Statistical Methods for Analysing Drug Sensitivity of Cancer Cells</td>
</tr>
<tr>
<td>17.15-17.18</td>
<td>Jos Hageman</td>
<td>Sample size determination for classification using metabolomics data</td>
</tr>
<tr>
<td>17.18-17.21</td>
<td>Rianne van Binsbergen</td>
<td>Accuracy of imputation and prediction in dairy cattle using whole-genome sequence data</td>
</tr>
<tr>
<td>17.21-17.24</td>
<td>Sabine Schnabel</td>
<td>Expectile smoothing for big data and visualization of linkage disequilibrium decay</td>
</tr>
<tr>
<td>17.24-17.27</td>
<td>Wessel van Wieringen</td>
<td>Ridge estimation of inverse covariance matrices from high-dimensional data</td>
</tr>
</tbody>
</table>

17.30-20.00  **Poster session, Drinks & Bites**
### Tuesday 21st April

#### 9.00-11.00 Morning session I

**Invited II, Integrated Population Modelling**

- **9.00-9.40** Marie-Pierre Etienne, Michael Schaub
  - Building robust abundance indices combining commercial data and scientific survey
- **9.40-10.20** Takis Besbeas
  - Differential contribution of vital rate synchrony on population synchrony in barn swallows
- **10.20-11.00** R. G. Besbeas
  - Recent developments in Integrated Population Modelling

**Contributed C5 (Bayesian Statistics)**

- **9.00-9.20** Cajo ter Braak
  - Bayesian P-spline smoothing using INLA
- **9.20-9.40** Philippe Lambert
  - A dynamic model for the incidence rate of ebola in Sierra Leone
- **9.40-10.00** Robert Goudie
  - Bayesian modelling for evidence synthesis via Bayesian melding
- **10.00-10.20** Maria Gheorghe
  - Using Bayesian Networks (BNs) for predicting patient-reported outcome measures
- **10.20-10.40** Pablo Verde
  - Comparing diagnostic test in meta-analysis: a hierarchical meta-regression approach

**Contributed C6 (Genetics and Omics)**

- **11.30-11.50** Monika Jelizarow
  - Analyzing sets of SNPs via SKAT versus global test: methodical overlaps and their practical consequences
- **11.50-12.10** Ron Wehrens
  - Fast Parametric Warping of Peaks in Metabolomics Data
- **12.10-12.30** Mar Rodriguez Girondo
  - Augmented prediction in Omics applications

**Contributed C7 (Statistical Modelling)**

- **11.30-11.50** Chamberlain Mbab
  - A demonstration of overfitting and model selection in a unique setting of two independent data sets.
- **11.50-12.10** Martin Boer
  - Fast P-splines for very long data series
- **12.10-12.30** Sjoukje Vandenberghe
  - Mediation analysis of randomised experiments

**Contributed C8 (Survival, prediction, joint models)**

- **11.30-11.50** Eleni Andrinopoulou
  - Bayesian variable selection with joint modeling of a longitudinal and survival outcome assuming different association structures
- **11.50-12.10** Loic Ferrer
  - Joint modelling of longitudinal and multi-state processes: an application in patients with localized prostate cancer
- **12.10-12.30** Anais Rouanet
  - Joint latent class model for longitudinal data and interval-censored semi-competitive events: Application to Alzheimer’s disease

#### 11.00-11.30 Break

#### 11.30-12.30 Morning session II

**Contributed C6 (Genetics and Omics)**

- **11.30-11.50** Monika Jelizarow
  - Analyzing sets of SNPs via SKAT versus global test: methodical overlaps and their practical consequences
- **11.50-12.10** Ron Wehrens
  - Fast Parametric Warping of Peaks in Metabolomics Data
- **12.10-12.30** Mar Rodriguez Girondo
  - Augmented prediction in Omics applications

**Contributed C7 (Statistical Modelling)**

- **11.30-11.50** Chamberlain Mbab
  - A demonstration of overfitting and model selection in a unique setting of two independent data sets.
- **11.50-12.10** Martin Boer
  - Fast P-splines for very long data series
- **12.10-12.30** Sjoukje Vandenberghe
  - Mediation analysis of randomised experiments

**Contributed C8 (Survival, prediction, joint models)**

- **11.30-11.50** Eleni Andrinopoulou
  - Bayesian variable selection with joint modeling of a longitudinal and survival outcome assuming different association structures
- **11.50-12.10** Loic Ferrer
  - Joint modelling of longitudinal and multi-state processes: an application in patients with localized prostate cancer
- **12.10-12.30** Anais Rouanet
  - Joint latent class model for longitudinal data and interval-censored semi-competitive events: Application to Alzheimer’s disease

#### 12.30-13.30 Lunch

#### 13.30-15.30 Afternoon session I

**Invited III, Confounder Modelling and Selection**

- **13.30-14.10** Tom Heskes
  - Robust causal discovery in the presence of latent variables
- **14.10-14.50** Ingeborg Warnbaun
  - Model misspecification and bias for inverse probability weighting and doubly robust estimators
- **14.50-15.30** Karel Vermeulen
  - Bias-Reduced Doubly Robust Estimation

**Contributed C9 (Survival, prediction, joint models)**

- **13.30-13.50** Birgit Witte
  - Estimating the type-specific cumulative incidence of cervical precancer in a setting with competing risks and masked event types
- **13.50-14.10** Jan De Neve
  - A unified effect measure for semiparametric linear transformation models
- **14.10-14.30** Roel Verbelen
  - Multivariate mixtures of Erlangs for density estimation under censoring and truncation
- **14.30-14.50** Theodor Balan
  - Event-based Ascertainment of Recurrent Events Data
- **14.50-15.10** Hein Putter
  - Dynamic frailty models based on compound birth-death processes
- **15.10-15.30** Leyla Azarang
  - Direct modeling of regression effects for transition probabilities in the non-Markov illness-death model
15:30-16:00  **Break**

16.00-17.20  **Afternoon session II**

**Contributed C10 (Ecology) (Red room)**

16.00-16.20  Natoya Jourdain  Abundance estimation using integrated modelling of unmarked animal data from camera traps.

16.20-16.40  Rianne Jacobs  Estimation of $P(X>Y)$ for normal distributions in the context of probabilistic environmental risk assessment for nanoparticles

16.40-17.00  Magne Aldrin  Modelling the sea lice population at a salmon farm

17.00-17.20  Diana Cole  Parameter Redundancy in Integrated Population Models

**Contributed C11 (Survival, prediction, joint models) (Yellow room)**

16.00-16.20  Mia Klinten Grand  Dynamic prediction by direct binomial regression

16.20-16.40  Jammbe Musoro  Dynamic prediction of mortality amongst patients in intensive care using the sequential organ failure assessment (SOFA) score

16.40-17.00  Nan van Geloven  Time dependent discriminative ability of dynamic prediction models with discrete survival times

**Contributed C12 (Longitudinal data & Causality) (Blue room)**

16.00-16.20  Nicole Erler  Multiple imputation for incomplete predictors and complex outcomes: mice vs. sequential bayesian imputation

16.40-17.00  Anna Ivanova  Pseudo-Likelihood Approach for Large and Complex Ordinal Data

17.00-17.20  Susan Bryan  Prediction of glaucoma progression using a high-dimensional two-stage Bayesian hierarchical model

19.00-23:00  **Diner**

The conference dinner will take place on Tuesday at Landmark Wijnfort, Bemmelsedijk 4, 6663 KZ Lent, near Nijmegen

T:  +31 (0)24 3231746, E: info@landmark.nl, W: www.landmark.nl/landmark-wijnfort/
## Wednesday 22nd April

**9.30-10.30**  
**Keynote (Lindenberg room)**  
Sylvia Richardson  
Analysis of complex phenotypes in genomics

**10.30-11.00**  
**Break**

**11.00-12.20**  
**Morning session II**

### Contributed C13 (Ecology) (Red room)

<table>
<thead>
<tr>
<th>Time</th>
<th>Presenter</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.00-11.20</td>
<td>Fred van Eeuwijk</td>
<td>P-spline models for agricultural field trials</td>
</tr>
<tr>
<td>11.20-11.40</td>
<td>Julie Aubert</td>
<td>Latent Block Model for ecological abundance data</td>
</tr>
<tr>
<td>11.40-12.00</td>
<td>Nadia Vendrig</td>
<td>Ultrasonic vocalizing behaviour in socially housed rats: a composite-link function approach</td>
</tr>
<tr>
<td>12.00-12.20</td>
<td>Paul Eilers</td>
<td>Capture-recapture and the penalized composite link model</td>
</tr>
</tbody>
</table>

### Contributed C14 (Longitudinal data & Causality) (Yellow room)

<table>
<thead>
<tr>
<th>Time</th>
<th>Presenter</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.00-11.20</td>
<td>Simon Thompson</td>
<td>Using pleiotropic genetic variants to simultaneously estimate multiple causal effects: a multivariable extension to Mendelian randomization</td>
</tr>
<tr>
<td>11.20-11.40</td>
<td>Ernst Wit</td>
<td>Estimating non-linear causal effects from observational data</td>
</tr>
<tr>
<td>11.40-12.00</td>
<td>Daniel Commenges</td>
<td>The dynamic approach to causal reasoning in ageing studies</td>
</tr>
</tbody>
</table>

### Contributed C15 (Statistical Modelling) (Blue room)

<table>
<thead>
<tr>
<th>Time</th>
<th>Presenter</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.00-11.20</td>
<td>Alan Welsh</td>
<td>Fitting linear mixed models with and without centering the explanatory variables</td>
</tr>
<tr>
<td>11.20-11.40</td>
<td>Aldo Solari</td>
<td>Variable selection and deselection in regression models, with confidence</td>
</tr>
<tr>
<td>11.40-12.00</td>
<td>Jesse Hemerik</td>
<td>Exact testing with random permutations</td>
</tr>
<tr>
<td>12.00-12.20</td>
<td>Yacin Yavuz</td>
<td>Count data analysis in nutrition clinical trials</td>
</tr>
</tbody>
</table>

**12.20-12.40**  
**Closing (Lindenberg room)**

**12.40-13.30**  
**Lunch**

### Satellite meeting, Inferring dynamic genetic networks  
(organised by Fentaw Abegaz and Ernst Wit, on the occasion of the Dutch Biometric Award)

<table>
<thead>
<tr>
<th>Time</th>
<th>Presenter</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>13.30-14.00</td>
<td>Fentaw Abegaz</td>
<td>Chain graphical models for inferring dynamic networks</td>
</tr>
<tr>
<td>14.00-14.30</td>
<td>Ines Wilms</td>
<td>Sparse vector autoregressive models with an application in marketing</td>
</tr>
<tr>
<td>14.30-15.00</td>
<td>Gerda Claeskens</td>
<td>A focused selection of graphical models in fMRI connectivity studies</td>
</tr>
<tr>
<td>15.00-15.20</td>
<td>Break</td>
<td></td>
</tr>
<tr>
<td>15.20-15.50</td>
<td>Dirk Husmeier</td>
<td>Controversy in mechanistic modelling of biopathways with Gaussian processes</td>
</tr>
<tr>
<td>15.50-16.20</td>
<td>Marco Gregorzyk</td>
<td>Realistic network reconstruction methods for the Timing Metabolism (TiMet) research project</td>
</tr>
<tr>
<td>16.20-16.50</td>
<td>Michael Eichler</td>
<td>Graphical models for causal inference from multivariate time series</td>
</tr>
<tr>
<td>16.50-17.00</td>
<td>Closing</td>
<td></td>
</tr>
<tr>
<td>17.00-17.45</td>
<td>Drinks</td>
<td></td>
</tr>
</tbody>
</table>
Keynote
In this talk, I will discuss some of our recent work in the area of statistical genomics, aimed at improving our biological understanding of complex phenotypes and disease process. Multiple layers of molecular phenomena are involved in the path between genes and disease. To better exploit the structure of the rich biological data that is currently being collected, statistical models linking the different layers of genomics structures are constructed. Generic Bayesian hierarchical model building and variable selection algorithms have been tailored for this purpose. In this talk, I will discuss application of these approaches for identifying prognostic tumour markers, detecting polygenic association with multiple phenotypes, modelling expression data, and analysing ontology data. The analytical strategy will be illustrated on case studies.
Invited Session 1: Dynamic prediction
Landmarking was originally introduced as a way of dealing with the problem of immortal
time bias in the context of time-dependent covariates in survival analysis. It has later been
proposed as a method of obtaining dynamic predictions for a survival outcome without
the need of constructing comprehensive models for the stochastic behaviour of intermedi-
ate events or longitudinal measurements in relation with the survival outcome of interest.
The aim of this talk is to show how landmarking can be used in conjunction with another
popular method in survival analysis, pseudo-observations, to construct dynamic regres-
sion models for non-standard outcomes. Two illustration of the use of dynamic pseudo-
observations will be given. The first is in a competing risks context, where interest is in
estimating the conditional cumulative incidence of a given cause, given covariates. In the
other application, interest is in direct regression models for expected residual healthy life
in an illness-death multi-state model.
Decision making in medicine has become increasingly complex for patients and practitioners. This has resulted from factors such as the shift away from physician authority toward shared decision making, unfiltered information on the Internet, new technology providing additional data, numerous treatment options with associated risks and benefits, and results from new clinical studies. Within this context medical screening procedures are routinely performed for several diseases. In this work we are interested in optimizing screening intervals for symptomatic patients that are followed-up prospectively, possibly after a medical procedure. Our methodological developments are based on the framework of joint models for longitudinal and time-to-event data. More specifically, our aim in this work is twofold: First, to appropriately select the joint model to use at time t, the time point the subject of interest is still event-free, and second, based on this model to select the optimal time point \( u > t \) to plan the next measurement. We measure optimality by the amount of information gained (using the Kullback-Leibler divergence and the expected cross-entropy) for the survival process given the history of the subject that includes both baseline information and his/her accumulated longitudinal measurements. The motivation for this research comes from a study conducted by the Department of Cardio-Thoracic Surgery of the Erasmus Medical Center in the Netherlands. This study includes 285 patients who received a human tissue valve in the aortic position. A major disadvantage of using human tissue is their susceptibility to degeneration and the concomitant need for re-interventions. The durability of a cryopreserved aortic allograft is age-dependent, leading to a high lifetime risk of re-operation, especially for young patients. It is therefore of great interest for cardiologists and cardio-thoracic surgeons to plan in the best possible manner echocardiographic assessments that will inform them about the future prospect of a patient with a human tissue valve in order to optimize medical care, carefully plan re-operation and minimize valve-related morbidity and mortality.

Dimitris Rizopoulos
Dept. Biostatistics, Erasmus MC, Rotterdam, Netherlands
e-mail: d.rizopoulos@erasusmc.nl
On methods for evaluating dynamic predictions

Paul Blanche

Boosted by the growing interest in personalized medicine, there is currently an increasing demand for dynamic predictions of patient prognosis. By dynamic predictions we mean predictions of a clinical event which are updated as soon as the time of making predictions changes. As an example, in this talk we focus on the case for which the predictions are based on the entire history of a marker trajectory at the time of making predictions. Such dynamic predictions are typically computed from joint-modeling or landmarking approaches. We focus on statistical methods for quantifying the predictive accuracy of such dynamic predictions. Different approaches are presented, including those based on ROC curves, prediction error curves and R2-type curves. We discuss how they evaluate the calibration and discrimination of the predictions, which are two essential concepts to evaluate the usefulness of any clinical prediction model. To deal with censoring, non-parametric inverse probability weighting estimators are suggested. Large sample results are provided. They enable the computation of confidence regions for pointwise and simultaneous inference. The ideas are motivated and illustrated by the analysis of large cohort data. The aim is to build and evaluate models of absolute risk of dementia.
Invited Session 2: Integrated Population Modelling
Building robust abundance indices combining commercial data and scientific survey

Marie-Pierre Etienne, Robyn Forrest, Jean-Baptiste Lecomte and Eric Parent

The quality of a stock assessment strongly depends on the quality of the underlying fitted population dynamic model used to investigate the consequence of different harvest rules. The estimation and even the selection of the appropriate model is mainly driven by the abundance indices considered. Those indices are derived either from marine scientific surveys or obtained with ad-hoc normalisation of the commercial catch, called Catch Per Unit of Effort (CPUE). Data sampled during marine scientific surveys and CPUE often present a high proportion of zeros with, possibly skewed, positive continuous values. In addition for commercial fisheries, the sampling process and the latent biomass process are stochastically dependent since professionals - conversely to scientists - use to concentrate in areas that are thought likely to yield high fish abundance. The spatial locations of the commercial catch provide information on the spatial biomass repartition. Furthermore, because of the spatial structure of the population and of the sampling process, a nave abundance indice produces biased estimates of the trends. The spatial aspect of the sampling process has to be accounted for. But zero inflated continuous data and preferential sampling prevent from using standard geostatistical methods. In addition, the fishermen data are cheap but massive and require to use model that can accommodate big data sets. We develop a model that addresses the two limiting specific features, give a general expression for the likelihood function, and discuss how inference can be performed, at least approximately, using advanced Monte Carlo methods. We finally describe a possible application of such a model to bottom fish data from the continental shelf of the coast of British Columbia, Canada collected both by marine scientists (Ocean and Fisheries, Canada) and by fishermen.

Marie-Pierre Etienne
AgroParisTech, Paris, France
e-mail: marie.etienne@agroapristech.fr
Differential contribution of vital rate synchrony on population synchrony in barn swallows

Michael Schaub, Johann von Hirschheydt and Martin U. Grüebler

Populations of many species show synchronous dynamics, mostly caused by spatial autocorrelation of the environment or dispersal. Often it can only be speculated about the demographic reasons of population synchrony, because demographic parameters are unknown. Here we develop a novel hierarchical integrated population model to analyze demographic data of barn swallows (Hirundo rustica) from 9 sites in Switzerland over 7 years. We decomposed the spatio-temporal variation of the demographic rates (apparent survival, components of productivity, immigration) into spatial and temporal components using random effects. This allowed the estimation of synchrony of the demographic parameters and of the population growth rate. The barn swallow populations fluctuated synchronously, but less so as most demographic parameters. The highest synchrony showed the probability of double brooding, while fledging success was highly asynchronous and the other demographic rates achieved intermediate levels of synchrony. Population growth was most sensitive to changes in immigration and adult apparent survival, and both of them contributed to the observed temporal variation of population growth rate. Using a simulation model we show that immigration and apparent survival of juveniles and adults were able to induce population synchrony, but not components of reproductive success due to their low population growth rate sensitivity. Since immigration and partially also apparent survival are linked with dispersal, our results indicate a prominent role of dispersal for population synchrony in barn swallows. The outlined approach seems promising for achieving a mechanistic understanding of population synchrony.
Recent developments in Integrated Population Modelling

Takis Besbeas

Integrated Population Modelling (IPM) has become an increasingly familiar technique for analysing wild animal data at the individual and population-level simultaneously. The method is based on a state-space model (SSM) for population size that connects the data sources together with parameters estimated by maximum-likelihood or within a more general Bayesian framework. We provide a general review of the work, with a focus on recent developments with respect to, for example, repeating the population survey to increase information from the monitoring and checking goodness-of-fit. We also build on the fine line of distinction between SSMs and hidden Markov models (HMMs) and present on-going work on fitting IPMs using HMM methodology. The methods of the talk are illustrated with real and simulated avian data.
Invited Session 3: Confounder Modelling and Selection
Robust causal discovery in the presence of latent variables

Tom Heskes and Tom Claassen

Although in theory constraint-based algorithms for causal discovery can handle latent variables, in practice they tend to be very sensitive to incorrectly estimated conditional (in)dependencies. In this talk, I will sketch our modular, probabilistic approach towards more robust causal discovery. The underlying idea is to derive local causal statements from so-called minimal conditional dependencies and independencies, that are then combined using straightforward logical deduction. Bayesian methods can be used to estimate the reliability of each individual statement and the final outcome. I will illustrate our approach on a data set related to ADHD and will discuss its potential for causal discovery from big data.
Model misspecification and bias for inverse probability weighting and doubly robust estimators

Ingeborg Waernbaum

Function of covariates, such as the propensity score, play an important role for confounder control when estimating an average causal effect. Different classes of estimators of a causal effect rely on model assumptions of, e.g., outcome regression or propensity score models in different ways. For a class of estimators that condition on a function of the covariates the average causal effect can be consistently estimated if the limit of the estimated covariate function converges in probability to a function that is sufficient for the assumption of no unmeasured confounding to hold. IPW estimators rely on consistent estimation of the propensity score. In this paper we study and compare a generic form of the bias of IPW and DR estimators under misspecifications of the propensity score and outcome regression models. We derive conditions under which the bias of a DR estimator is larger than the bias of an IPW estimator when both the outcome model and the propensity score model are misspecified.

Ingeborg Waernbaum
University Umeå, Umeå, Sweden
e-mail: ingeborg.waernbaum@stat.umu.se
Bias-Reduced Doubly Robust Estimation

Karel Vermeulen and Stijn Vansteelandt

Over the past decade, doubly robust estimators have been proposed for a variety of target parameters in causal inference. These are asymptotically unbiased when either an outcome regression model or a propensity score model is correctly specified, regardless of which. While theoretically appealing, doubly robust estimators have been the subject of recent debate. The reason is that model misspecification is likely to affect all working models in practice, and thus the very premise that at least one of both working models is correctly specified, lives on shaky grounds. Moreover, the performance of doubly robust estimators can be sensitive to the choice of estimators used for fitting the working models, and can sometimes be worse than that of competing estimators that do not enjoy the double protection property. In this talk, I will show that, interestingly, some doubly robust estimators partially retain their robustness properties even under misspecification of both working models. In particular, I will propose a simple and generic estimation principle for the nuisance parameters indexing each of the working models, which we call bias-reduced doubly robust estimation (Vermeulen and Vansteelandt, 2015). It is designed to improve the performance of the doubly robust estimator of interest, relative to the default use of maximum likelihood estimators for the nuisance parameters by locally minimising the squared first-order asymptotic bias of the double-robust estimator under misspecification of both working models. We discuss the basic proposal, which is based on parametric models for the nuisance models, as well as extensions that employ machine learning algorithms. Simulation studies confirm the desirable finite-sample performance of the proposed estimators relative to other proposals.

Contributed Session C1: Genetics and Omics
Testing for splice-QTL effects: finding then SNPs and exons in the haystack

Renee de Menezes and Marianne Jonker

Each gene is formed by a number of exons, which may vary between one and a couple of hundreds. Different transcripts from the gene can be produced if, during transcription, different subsets of exons are used; this process is called alternative splicing. Since different transcripts may result in differential protein expression, it is important to understand this regulatory mechanism for unravelling the biological process underlying, for instance, the development of cancer. Here we are mainly interested in finding genes with alternative splicing events that are associated with SNP-genotypes within or near the gene (called splice-QTL). Subsequently, we want to identify SNPs that drive the splicing events. For this, we used 373 samples from the HapMap project, for which whole-genome exon expression and SNP genotypes were available. Per gene, there can be up to a couple of hundreds exons, and typically a much larger number of SNPs. We first perform a score test to find genes which exons have expression associated with the SNPs. This score test represents the SNP effects as random effects that depend both on the exon as well as on the SNP. After selecting splice-QTL genes, driving SNPs are selected using simultaneous variable selection via a penalized regression method. By using data from the GEUVADIS and 1000Genomes projects, our test can find splice-QTL effects even when different SNPs display association with different exons. This power to find a wide range of effect types makes our method a fundamental tool to further unravel the workings of SNP-driven exon regulation.

Joint work with Marianne Jonker (VUmc) and Jakub Pecanka (Radboud umc)

Renee de Menezes
VU University Medical center, Amsterdam, the Netherlands
e-mail: r.xavierdemenezes@vumc.nl
Integration is a central theme in multiplatform data analysis. Multiplatform analysis represents the integrative analysis of (biological) systems measured with different technologies. Differences can arise in applying different techniques, or even fundamentally measuring different biological particles. Combining different datasets generated by different technologies or platforms can provide more information and may yield a better understanding of the underlying system that is measured. Key in this work is assessing to what extent two different technologies measure the same underlying omic. The two-way orthogonal partial least squares method (O2PLS, Trygg, 2003) handles multiplatform modelling in a symmetric way. O2PLS has similar disadvantages as PLS. A lack of a probabilistic foundation makes interpreting the results of O2PLS difficult. Moreover, hypotheses are in many cases not well defined. Complex estimation procedures make it difficult to identify potential estimation problems. Also the assumptions may not be sensible if used as population model. In this work we propose a novel probabilistic approach (PO2PLS). It provides mathematical and statistical clarity about the O2PLS model, as well as aiding in the interpretation in large epidemiological studies. We will expose and discuss potential drawbacks of O2PLS that are hard to detect without this probabilistic framework. A new extension is the development of an information score based on PO2PLS, indicating how much or little two data sets are related. Simulation studies will be presented for assessing the PO2PLS performance and it will be apply to glycomics data. The glycans were measured on N=681 subjects with two different platforms, LC-MS (p=25) and MALDI-FTICR (q=60). These techniques measure the same system, but in a different way. We will quantitatively investigate the similarity of the two data sets. The contribution of all glycans will be visualised. This will give insight and more efficient use of glycomics when complex models need to be fitted.
Reconstruction of genome ancestry blocks in multiparental populations

Chaozhi Zheng, Martin P. Boer and Fred A. van Eeuwijk

We present a hidden Markov model framework called RABBIT for reconstructing genome ancestry blocks from single nucleotide polymorphism (SNP) array data, a required step for quantitative trait locus (QTL) mapping. The framework can be applied to a wide range of mapping populations such as the mouse Collaborative Cross (CC) and the Arabidopsis multiparent advanced generation intercross (MAGIC) for both autosomes and X chromosomes if they exist. The model underlying RABBIT accounts for the joint pattern of recombination breakpoints between two homologous chromosomes, and missing data and allelic typing errors in the genotype data of both sampled individuals and founders. Studies on simulated and real data of the CC and the MAGIC demonstrate that RABBIT is more robust and accurate in reconstructing recombination bin maps than some commonly used methods.

Chaozhi Zheng
Wageningen UR, Wageningen, Netherland
e-mail: chaozhi.zheng@wur.nl
GLMMs for nucleic acid concentration estimation in digital droplet PCR

Matthijs Vynck, Olivier Thas and Jo Vandesompele

Digital droplet polymerase chain reaction (ddPCR) is a recently developed alternative for nucleic acids quantification that allows for more accurate and precise estimation of nucleic acids concentrations. By splitting up a sample into about 14000 droplets using microfluidic technology and subsequently counting the numbers of negative (no nucleic acid was initially present) and positive (nucleic acid was initially present) droplets, the concentration of the nucleic acid can be determined by using a Poisson distribution. Often, the procedure is repeated (technical replication) to further increase the accuracy and precision of the concentration estimate. To estimate a copy number variation (CNV), the quantification process is conducted in parallel: the presence or absence for two nucleic acids is measured simultaneously by using two different fluorescent dyes, thus resulting in two concentration estimates. Taking a ratio of these estimates finally results in an estimate for the CNV. We show how the concentration estimation follows from using the generalized linear mixed model (GLMM) framework and how it extends to the CNV case. Furthermore, the GLMM framework also allows for estimation of the concentration and CNV by making use of the technical replicates of the samples through introduction of a random effect. We illustrate the importance of this increased accuracy and precision for applications in a medical diagnostic setting.
Measuring telomere lengths using DNA sequencing data: Discovering even more layers of heterogeneity in Prostate Cancer

Andy Lynch, J Henry Farmery, Charles E Massie and Mike L Smith

Telomeres (the ends of chromosomes that protect the genome from the shortening that occurs with DNA replication) play an integral part in tumourigenesis either through the need to overcome the obstacle of telomere-shortening (which should trigger an end to replication) or through the tumour-driving structural mutations that can arise in the genome if replication somehow occurs when the telomeres are too short to provide protection. Next generation sequencing of DNA does not ignore these regions, although analyses of these data often do. Due to the importance of telomeres in cancer, it is natural that we should wish to estimate telomere lengths from the abundance of cancer sequencing data now in existence. Yet, while there have been a handful of efforts to quantify telomeres from sequencing data, these do not adequately address the problem for tumour samples. In brief, existing methods count the number of sequencing reads that look telomeric, and cannot translate this to an estimate of telomere length unless the number of telomeres can be pre-specified which is often not the case for tumours due to frequent chromosomal copy number changes. Additionally, there is no accounting for regions of the genome that look, but are not, telomeric and most do not account for known sequencing biases such as that driven by the proportion of G and C bases present. We present Telomerecat: a tool implementing a new approach to the estimation of the mean telomere length from tumour sequencing data. Telomerecat combines statistical inference with an understanding of sequencing technologies in order to overcome the limitations of existing approaches. Using data from the International Cancer Genome Consortium UK Prostate Cancer initiative we show that there exists heterogeneity of telomere length within cancerous prostates, and that the estimated lengths associate with genetic, epigenetic and transcriptomic activity in the cancer.
Contributed Session C2: Genetics and Omics
Directed cyclic mixed graph modeling for network integration of high-dimensional genomic data

Carel F.W. Peeters, Wessel N. van Wieringen and Mark A. van de Wiel

Many graphical modeling efforts set out to infer an undirected network or to map an undirected graph to a recursive (acyclic) functional model. Many data-generating processes, however, are subject to cycles or feedback relations. The desire is to capture, through graphical modeling, such dynamics systems in equilibrium. Our methodological objective is then to infer (sparse) functional graph patterns from Gaussian precision matrices structured according to nonrecursive systems of simultaneous equations in which (a) certain subsets of variables may enjoy differential treatment, and (b) reciprocal relations and cycles are allowed. The resulting graphs belong to the class of directed cyclic mixed graphs. The framework we propose intends to be attractive in terms of scalability. Our applied goal is to bring the resulting framework to bear on integrative graphical modeling of oncogenic data from multiplegenomic platforms. The genomic platforms we consider regard messenger Ribonucleic Acid (mRNA) and microRNA expression data. The framework thus encompasses observational steady-state genomic data to functional analysis of the cyclic kind and, through the combination of different types of genomic data, supports more robust identification of deregulated cancer-pathways.

Carel F.W. Peeters
VU University Medical center, Amsterdam, the Netherlands
e-mail: cf.peeters@vumc.nl
tigaR: Integrative significance analysis of temporal differential gene expression induced by genomic abnormalities

Viktorian Miok, Saskia M Wilting, Mark A van de Wiel, Annelieke Jaspers, Paula I van Noort, Ruud H Brakenhoff, Peter JF Snijders, Renske DM Steenbergen and Wessel N van Wieringen

To determine which changes in the host cell genome are crucial for cervical carcinogenesis, a longitudinal in vitro model system of HPV-transformed keratinocytes was profiled in a genome-wide manner. Four cell lines affected with either HPV16 or HPV18 were assayed at 8 sequential time points for gene expression (mRNA) and gene copy number (DNA) using high-resolution microarrays. Available methods for temporal differential expression analysis are not designed for integrative genomic studies. Here, we present a method that allows for the identification of differential gene expression associated with DNA copy number changes over time. The temporal variation in gene expression is described by a generalized linear mixed model employing low-rank thin-plate splines. Model parameters are estimated with an empirical Bayes procedure, which exploits integrated nested Laplace approximation for fast computation. Iteratively, posteriors of hyperparameters and model parameters are estimated. The empirical Bayes procedure shrinks multiple dispersion-related parameters. Shrinkage leads to more stable estimates of the model parameters, better control of false positives and improvement of reproducibility. In addition, to make estimates of the DNA copy number more stable, model parameters are also estimated in a multivariate way using triplets of features, imposing a spatial prior for the copy number effect. With the proposed method for analysis of time-course multilevel molecular data, more profound insight may be gained through the identification of temporal differential expression induced by DNA copy number abnormalities. In particular, in the analysis of an integrative oncogenomics study with a time-course set-up our method finds genes previously reported to be involved in cervical carcinogenesis. Furthermore, the proposed method yields improvements in sensitivity, specificity and reproducibility compared to existing methods. Finally, the proposed method is able to handle count (RNAseq) data from time course experiments as is shown on a real data set.

Viktorian Miok
VU University Medical center, Amsterdam, the Netherlands
e-mail: v.miok@vumc.nl
Statistical methods to analyze repeated measurements of overdispersed categorical data: an application in longitudinal microbiome data

Ivonne Martin, Roula Tsonaka and Jeanine Houwing-Duistermaat

Statistical modelling of clustered categorical data can be challenging. Our work is motivated by a longitudinal study on human gut microbiome measurements where multivariate count data distributed over a number of bacterial categories are available. Typically dimension reduction is obtained by considering the composition at phylum level (e.g. 6 categories in our dataset). In the motivating study, bacterial data for subjects from helminth endemic areas at two time points are collected and the question of interest is to assess the effect of infection status on the bacteria composition. The multinomial regression model can be used for analysis. However, the presence of the correlation within a study subject (overdispersion) and the clustering over time need to be modeled properly. To account for overdispersion, the Dirichlet-multinomial model (DMM) is typically used. Since Dirichlet is a conjugate distribution of the multinomial, fitting this model is straightforward compared to the multinomial logistic mixed model (MLMM). However, the linear predictor models the Dirichlet parameter and not the mean which may hamper interpretation of the parameters. Other drawbacks are that, the correlations between the bacteria categories are forced to be all negative and extension to a multilevel data is not straightforward. We propose to extend the MLMM and use the combined model (CM), in which we incorporate two sets of random effects: a Dirichlet distributed random effect at the mean level to accommodate the overdispersion and a set of normally distributed random effects included in the linear predictor to model the association between the repeated measurements. We study the correlation between the various bacteria categories at cross sectional setting for the three models (DM, MLMM, CM) and evaluate the performance of our new method (CM) via simulations. Finally, we exemplify our proposed method on the bacterial data and estimate the effect of infection status over time in longitudinal setting.

Ivonne Martin
Leiden University Medical Centre, Leiden, The Netherlands
e-mail: I.Martin@lumc.nl
Contributed Session C3: Statistical Modelling
Estimating prevalences using medication data and random forests

L. Slobbe, K. Fussenich, A. Wong, C. van den Brink, H. Boshuizen, J.J. Polder and H. van Oers

In the Netherlands, for many diseases prevalence estimates are based on GP registrations. In an attempt to improve these estimates, diagnosis data from the GP registrations is linked to health insurers medication data. The advantage is that medication data is available for the whole Dutch population, whereas the diagnosis is unknown for most people. The best predicting medication combination for a specific diagnosis is found by using a random forest model. Each person is then allocated a probability of having a diagnosis. Summing up the probabilities gives a prevalence for the desired population. The model fit is evaluated using a Hosmer-Lemeshow test, and varies strongly by disease. For many diseases, such as diabetes, COPD, depression or MS, the model gives a good fit. Even though this method seems to provide good estimates for a number of diseases, there can be limitations when probabilities on diseases vary strongly by GP.

K. Fussenich
UMCG, Grongingen, The Netherlands
email: koen.fussenich@rivm.nl
Analysis for single patient studies

Paul Garthwaite, Fadlalla Elfadaly and John Crawford

Statistics rarely tries to draw conclusions from a single individual. However, in neuropsychology an individual with brain injury may display unusual combinations of abilities that give insight into the architecture of the brain, perhaps showing that two tasks that appear similar are, in fact, performed by different parts of the brain as the brain injury has impaired performance on one task but not the other. Clinical neuropsychologists also need to evaluate whether a patient has a deficit on some specified task, asking the question: How extreme are the patients scores compared with the scores that would be obtained in the general population? This talk addresses the task of estimating the proportion of the general population that would obtain a lower score (or more extreme score combination) than an individual obtained. The task seems simple but it is important to do it well because it arises so often. This leads to a surprising variety of statistical methods, including the use of non-central distributions and a means of determining the contribution of individual scores to a Mahalanobis distance. The latter helps explore features that make a persons profile unusual. Methods are implemented in user-friendly free software that may be run over the web without downloading it to ones computer. Data input is minimal: only sample means, standard deviations and correlations are required.

---

Paul Garthwaite
Open University, Milton Keynes, UK
e-mail: p.h.garthwaite@open.ac.uk
Reliability and agreement studies are of paramount importance in medical, behavioural and social sciences. They do contribute to the quality of studies by providing information about the amount of error inherent to any diagnosis, score or measurement. When raters classify independent units on an ordinal scale, weighted kappa coefficients are usually used to assess the reliability/agreement between or within raters. We propose a family of weighted kappa coefficients to quantify ordinal agreement between a pair of observers when they are considered as random. The case of random observers is frequently encountered in practice when for example interest lies in the pairwise agreement within a sample of observers (like during training sessions) or when a sample of subjects is rated by two observers, perhaps different for the subjects. We further develop a partial-Bayesian method, based on Generalized Linear Mixed Models, to evaluate the influence of predictors on these coefficients. Weighted agreement coefficients are directly related to a set of covariates through a multilevel non-linear model. This method naturally extends to account for multilevel data structures, very common in medical, behavioral and social sciences, where measurements are often obtained on persons nested in organizations, on various body parts or by repeated measurements over time. This work will help researchers to improve agreement levels between and within raters by first identifying factors influencing reliability/agreement coefficients and then by changing covariate values and improving the training of the raters whenever possible. This joint work with Pr. E. Lesaffre (KUL, Belgium), is supported by Netherlands Organisation for Scientific Research (grant 451-13-002).
Contributed Session C4: Survival, prediction, joint models
Summarising Censored Survival Data Using The Mean Residual Life Function

John Hinde, Alberto Alvarez-Iglesias, John Newell and Carl Scarrott

The mean residual life (MRL) function provides a clear and simple summary of the effect of a treatment or a risk factor in units of time, avoiding hazard ratios or probability scales that require careful interpretation. Estimation of the MRL is complicated by the upper tail of the survival distribution not being observed as, for example, patients may still be alive at the end of the follow up period. Various approaches have been developed to estimate the mean residual life in the presence of such right censoring. In this work, a novel semi-parametric method that combines existing non-parametric methods and an extreme value tail model is presented, where the limited sample information in the tail (prior to study termination) is used to estimate the upper tail behavior. This approach will be demonstrated with simulated and real-life examples.
Dementia is a major public health problem which seems to be increasing, mainly because of the increase of life expectancy. To try to reduce dementia burden, some intervention on dementia through its risk factors can be considered. However, it is important to assess the anticipated impact of the intervention on several health indicators. We propose a general framework to make projections for the prevalence of the disease, life expectancies with and without the disease, the age at onset and the life-long probability of the disease and evaluate the consequences of preventive interventions targeting risk factors in the whole elderly population on these various measures of the disease burden. The methodology takes into account the mortality trend over calendar time and age in both healthy and diseased subjects and the change in mortality due to the intervention.
A probabilistic record linkage model for survival data

Michel Hof and Koos Zwinderman

In medical research, record linkage is often used to combine baseline measurements (e.g. cross-sectional study data) with event data from external data sources (e.g. death registries). Unfortunately both datasets often lack a unique identifier that can distinguish record pairs that belong together (matches) from record pairs that do not (non-matches). In this situation, we could rely on partially identifying variables which have been registered in both data sources (e.g. gender or date of birth). However, the combined strength of the partially identifying variables is often not enough to perfectly distinguish matches from non-matches and bias is introduced in the analysis.

We propose a mixture model in which we combine a proportional hazards model with a probabilistic record linkage model. The matching indicator is treated as missing data. The computational effort required to evaluate the likelihood of our model depends on the matching structure imposed in the data. Without any structure, each record from both datasets could potentially match with all records from the other dataset. On the other hand, we could restrict each record from both data sources to have no or just one match in the other dataset. Generally, the more strict the matching structure, the more computational effort is required. Our model was developed using a N-1 matching structure; a baseline measurement could maximally match with one event time and an event time could match with multiple baseline measurements. Not only was this matching structure computationally friendly and thus applicable to large datasets, it also allowed us to re-write all the components from the mixture model in terms of survival probabilities. Simulations showed that our model gives accurate estimates in many different scenarios. Even with weak partially identifying variables, the parameter estimates were unbiased, had low MSEs, and excellent coverages of the 95

Michel Hof
UvA-AMC, Amsterdam, the Netherlands
e-mail: m.h.hof@amc.uva.nl
Contributed Session C5: Bayesian Statistics
Bayesian P-spline smoothing using INLA

Cajo J.F. ter Braak, María Xosé Rodríguez-Ivarez, Martin Boer, Paul Eilers and Havard Rue

P-splines are computational efficient smoothers, of which the working can easily be understood in terms of the degree of the B-splines and of the order of the difference penalty. The efficiency stems from the sparse basis representation of the B-spline. The sparseness is, however, lost by fitting the penalty parameters using mixed model software. Moreover, mixed models do not account for uncertainty in the variance components (and thus the penalty parameters) when generating confidence bounds for predictions. A Bayesian approach to smoothing easily allows to include the uncertainty. Integrated nested Laplace approximation (INLA) makes the Bayesian approach practical and the INLA R package already includes smoothing functions using piece-wise linear basis functions with peaks at the data points, supported by nice theory and links to geo-statistical modelling. We explore the possibility of fitting Bayesian P-splines using INLA thereby combining the advantages of the Bayesian approach using INLA with the flexibility and power of P-splines. In our approach sparseness is maintained. We are even able to go beyond generalized additive models without increasing the number of penalty parameters beyond the number of covariates. We wrote a small R library (BayesianPspline) that wraps up the approach. We illustrate our approach using an ecological example.
A dynamic model for the incidence rate of ebola in Sierra Leone

Philippe Lambert

It is thought that the ebola epidemic in Sierra Leone started end of May 2014 with a tribal healer claiming that she could cure people from ebola. Her patients were coming from Guinea where the epidemic made its first victims in December 2013. Women from Sierra Leone were infected at the occasion of her burial and contributed to the spread of the disease in the country. The Ministry of Health and Sanitation of Sierra Leone reports on the epidemic since then with statistics on the number of confirmed (by laboratory tests) cases starting in an irregular way early June and on a daily basis since mid-August. We extend the SEIR model (a compartmental model frequently used in epidemiology) to describe the evolution of the number of confirmed cases. Inference is made in a Bayesian framework to facilitate the quantification of uncertainty and to include prior information coming from individual follow-up of patients. The basic reproduction number, giving the expected number of secondary cases caused by an infected individual during the course of the disease, is monitored to evaluate the probability that the disease will die out.

This is joint work with Gianluca Frasso, Wilfried Bonou, Florence de Longueville and Pierre Ozer (Universit de Liege, Belgium).
Bayesian modelling for evidence synthesis via Bayesian melding

Robert Goudie, Anne Presanis, Lorenz Wernisch, Daniela De Angelis and David Lunn

We consider situations in which we wish to incorporate evidence from various sources within our statistical analysis, as in evidence synthesis. Separate statistical sub-models for each part of the evidence are typically available. Using sub-models reduces the dimensionality of the models that need to be fitted (thus simplifying computation) and can also make it easier to identify problematic areas in a model. However, it can be unclear how to transfer evidence” between the sub-models. In practice, evidence is typically transferred via a simple approximation. However, the form of the implied overall model may not be clear, and feedback between the sub-models is typically precluded. We propose a simple, general methodology for combining such sub-models into a single, joint statistical model. Our approach builds upon methods for combining the opinions of multiple experts in a prior elicitation exercise and upon Bayesian melding (Poole and Raftery, 2000). We also describe how inference for the joint model can be broken down into stages, in which only a single sub-model is considered at a time. In the first stage, we approximate the posterior distribution of a sub-model using Markov chain Monte Carlo (MCMC) and store the generated samples from the posterior distribution of each parameter of interest for use in the second stage. In the second stage, we use the samples generated in stage one to form ”proposal distributions” for use within a Metropolis-Hastings algorithm. We give an example of use of the methodology in evidence synthesis.”

Robert Goudie
MRC Biostatistics Unit, Cambridge, United Kingdom
e-mail: robert.goudie@mrc-bsu.cam.ac.uk
Using Bayesian Networks (BNs) for predicting patient-reported outcome measures

Maria Gheorghe, Anca Hanea, Melinde R.S. Boland, Maureen Rutten-van Mölken and Pieter H.M. van Baal

The use of patient-reported outcome measures (PROMs) such as EuroQoL-5D (EQ-5D) is mandatory for generating evidence to support reimbursement decisions. However, these measures are often not included in clinical studies, instead disease-specific questionnaires are collected. In these situations, it is accepted to predict or map PROMs from other disease-specific measures. For example, for mapping various disease-specific measures onto EQ-5D, various regression methods have been used including ordinary least squares (OLS), tobit, censored least absolute deviation (CLAD) and multinomial logit models (MNL). It is known that these methods have certain limitations such as predictive values that are outside of the domain, ceiling/floor effects and low prediction performance.

We propose using Bayesian Networks (BNs) for mapping disease-specific measures onto EQ-5D responses. BNs are probabilistic graphical models that explicitly model all relationships between the study variables without making any of the functional or distributional assumptions that exist in the regression-based methods. We used data collected from two randomized control trials, one for testing the model (N=5660 observations) and another one for validating the model (N=366 observations), for patients diagnosed with Chronic Obstructive Pulmonary Disease (COPD). Both of these trials include a disease-specific questionnaire for COPD patients and the EQ-5D questionnaire. We compared in-sample and out-of-sample prediction errors of the BNs model to current mapping methods: OLS, CLAD, Tobit and MNL.

Results: Our findings indicated that BNs consistently outperformed the other regression-based mapping models. For example, we found that overall out-of-sample errors such as mean absolute error and mean square error obtained with BNs mapping models were 0.078 and 0.015, respectively, and that these errors represented about half of the prediction errors obtained with regression-based methods. Compared to currently mapping methods in use, BNs provide a new flexible approach for mapping disease-specific questionnaires onto PROMs.

________________________________________

Maria Gheorghe
Erasmus University, Rotterdam, the Netherlands
e-mail: gheorghe@bmg.eur.nl
Comparing diagnostic test in meta-analysis: a hierarchical meta-regression approach

Pablo Verde

Meta-Analysis is a branch of statistical methods that allows researchers to combine scientific results from multiple pieces of evidence into a single analysis. Combining empirical evidence can be a challenge in practice; the data we wish to combine may have different grades of quality and information. As a consequence, special statistical techniques have to be applied. In this work we present a new meta-analysis approach to combine diagnostic results, when two diagnostic tests are applied to the same group of patients. The main problem in this type of meta-analysis is the lack of published information to make direct comparison between tests and to account for intra-study correlation of accuracy characteristics (e.g. sensitivities and specificities). We proposed a new Bayesian hierarchical meta-regression model, where the observed diagnostic results are considered as a realization of a latent process which allows direct comparison between tests. A new theoretical result establishing the relationship between observed and unobserved quantities is presented and used for model building. Variability between studies is modeled by using a mixture of multivariate scale normal distributions. This approach is illustrated with two systematic reviews: The first one investigates the diagnostic accuracy of doctors aided with decision tools (e.g. neural networks) compared with unaided doctors in patients with acute abdominal pain. The second one compares the diagnostic accuracy of positron emission tomography with computer tomography in the detection of lung cancer.
Contributed Session C6: Genetics and Omics
Analyzing sets of SNPs via SKAT versus global test: methodical overlaps and their practical consequences

Monika Jelizarow and Jelle J. Goeman

Among the major objectives of today's sequencing studies is to test predefined sets of single-nucleotide polymorphisms (SNPs) for association with some clinical parameter. Compared to conventional SNP-by-SNP analyses, such set-based analyses can both be more powerful and provide deeper biological insight into the phenomenon under study. One statistical method that has become a popular choice in this context is the Sequence Kernel Association Test, or briefly SKAT. Feasibility in high-dimensional data situations, the possibility of covariate adjustment and analytical p-value calculation render this test amenable to broad and efficient use in practice. In this talk we shall contrast SKAT with the well-known global test, which was originally developed for the analysis of gene expression data but which may likewise be employed for the analysis of SNP data, as numerous applications have illustrated. To begin with, we show that the SKAT statistic may in principle be seen as a reinvention of the global test statistic, and we point out the immediate practical consequences that arise from this. Subsequently, by means of simulations, we compare SKAT and global test in terms of type I error rate control, since they are based on different analytical approaches to obtain p-values. Finally, we discuss how our findings can be exploited to further improve the statistical analysis of SNP sets.
Fast Parametric Warping of Peaks in Metabolomics Data

Ron Wehrens, Tom G. Bloemberg and Paul H.C. Eilers

In many omics fields like proteomics and metabolomics, it is necessary to align the data so that features in different measurements actually mean the same things. This can be extremely complicated, and several tools have been developed to achieve this. Parametric Time Warping (PTW) has shown to be very efficient, and strikes a good balance between flexibility and power. In all general-purpose warping algorithms, alignment is performed on continuous profiles. Here we present a novel development in PTW allowing discrete features, consisting of lists of peak positions and intensities, to be mapped onto each other. The result is not only a speed increase of at least an order of magnitude, but in some cases also improvements in quality, since irrelevant variations in the original continuous signals no longer influence the warping. Moreover, the peaks on which the new warping methodology is based are generated anyway (by many different software packages) so that no additional effort is required. We will show the key ideas in this new approach and present a couple of examples from the area of metabolomics, highlighting the new possibilities. These include elements that hitherto were completely impractical, such as more time-intensive optimization routines than simply steepest-descent based optimizers in order to avoid local optima.

---

Ron Wehrens
Biometris, Wageningen UR, Wageningen, the Netherlands
e-mail: ron.wehrens@wur.nl
Augmented prediction in Omics applications

Mar Rodriguez-Girondo and Bart Mertens

A lot of attention has been devoted in the last years to accommodate single high-dimensional sources of molecular data (omics) in prediction models of health traits. During the last ten years, genomic and transcriptomics data have been widely used omic sources for prediction of health outcomes, alone or in combination with (low-dimensional) clinical covariates. Nowadays, other omic sources, such as proteomics, metabolomics, and glycomics are emerging as potentially interesting molecular sources for prediction of health-related traits. As a result, it is increasingly common to collect several omic measurements in the same set of individuals, and hence, new statistical challenges are emerging nowadays, namely, how to combine all these new information and to quantify the additional value of new molecular sources over previously established ones. Our motivating example illustrates these difficulties. We consider the Dietary, Lifestyle, and Genetic determinants of Obesity and Metabolic syndrome (DILGOM) study, sampled from the Helsinki area, in Finland. Data on serum metabonomes, genome-wide profiles of genetic and transcriptional variation from blood leukocytes are available, jointly with a large number of clinical and demographic factors. We are interested in investigating the role of each of the omic sources in the prediction of BMI and their respective additional predictive value with respect to each other. We propose a two-step procedure based on sequential double cross-validation prediction and regularized regression models, i.e., we consider the problem of combination of omic predictors in an ‘asymmetric’ way by sequentially assessing the augmented predictive ability of omic sources with respect to a given outcome of interest. Namely, we propose several performance indices to summarize the relation between the omic sources under study and a permutation test to formally assess the augmented predictive value of a second omic set of predictors over a primary omic source.
Contributed Session C7: Statistical Modelling
A demonstration of overfitting and model selection in a unique setting of two independent data sets.

Chamberlain Mbah, Olivier Thas, Jan De Neve, Hubert Thierens, Jenny Chang-Claude, Petra Seibold, Irmgard Helmbold, Sabine Behrens, Odilia Popanda, Peter Schmezer and Kim De Ruyck

Toxicity after radiotherapy remains a threat to the quality of life of breast cancer patients. Having access to good prediction models of toxicity is expected to improve decision making before and during radiotherapy. In this talk we demonstrate issues of overfitting and model selection in the context of binary prediction models and in the exceptional situation of having data from two independent cohorts. This allows us to use classical techniques on one dataset (e.g. cross-validation, lasso), and to evaluate the final model on the other cohort.

Chamberlain Mbah
University of Ghent, Ghent, Belgium
e-mail: chamberlain.mbah@ugent.be
Fast P-splines for very long data series

Martin Boer and Paul Eilers

P-splines combine regression on a B-spline basis with a difference penalty. Usually the number of B-splines is moderate, say less than 100, but in some applications we encounter very long data series and sparseness becomes crucial. An example is high-resolution X-Ray diffraction, delivering scans with many thousands of observations. When light smoothing is needed, the number of B-splines may be a thousand or more. For a given value of the penalty parameter, the B-spline coefficients can be calculated quite efficiently, because of the local character of the B-spline bases. In other words, minimizing the penalized sum of squares results in a sparse linear system of equations, and the corresponding coefficient matrix $M$ is banded. This system can be solved quite efficiently with standard algorithms, and the calculation time and computer memory needed increase linearly in the number of knots. The sparseness of $M$ is a nice property, but it can be destroyed in additional calculations, needed to optimize the penalty weight. For example, the standard equation used to calculate the effective dimension includes the inverse of $M$, which is a dense matrix. Another example is the transformation to a mixed model formulation. Existing transformations, such as the one using the singular value decomposition of the difference penalty matrix, destroy the local character of the B-spline basis. We present a new algorithm, in which the sparseness of the P-spline model is maintained in the mixed model formulation, by using automated differentiation and the special properties of B-splines. We show results for real and simulated data. We also explain how to maintain sparseness in when constructing the basis itself.
Mediation analysis of randomised experiments

Sjouke Vandenberghe and Stijn Vansteelandt

In this talk, we will make use of mediation analysis to decompose the effect of a randomised treatment on an end-of-study outcome into its indirect effect via a certain mediator and the remaining direct effect. This is motivated by a re-analysis of the EORTC 10994/BIG 1-00 randomised phase 3 trial, which aims to infer the indirect effect of taxane and anthracycline based chemotherapies on overall survival of advanced breast cancer patients, that is mediated by pathological complete response. We improve the efficiency of estimators of natural direct and indirect effects that were previously proposed by Tchetgen Tchetgen (2011) and Tchetgen Tchetgen and Schipitser (2012), by exploiting the fact that the data originate from a randomised experiment. The resulting estimators are less model-dependent than popular estimators based on the mediation formula: they do not demand correct specification of the model for the mediator, because they rely on the known randomisation probabilities. Results from a simulation analysis are shown, in which the proposed estimators are compared with competing estimators in terms of efficiency and bias in a variety of settings, for example under misspecification of the mediator model. Data analysis results will be discussed for the EORTC 10994/BIG 1-00 randomised phase 3 trial. This is joint work with Luc Duchateau (UGent), Achmad Efendi (EORTC) and Leen Slaets (EORTC).

Sjouke Vandenberghe
Ghent University, Ghent, Belgium
E-mail: sjouke.vandenberghe@ugent.be
Contributed Session C8: Survival, prediction, joint models
Bayesian variable selection with joint modeling of a longitudinal and survival outcome assuming different association structures

Eleni-Rosalina Andrinopoulou

The joint modeling of longitudinal and survival data has recently received a lot of attention. Several extensions of the standard joint model that consists of one longitudinal and one survival outcome have been proposed including among others the consideration of different association structures between the longitudinal and the survival outcomes. However, in general, relatively little attention has been given to the selection of the most appropriate functional form to link the two outcomes. In common practice, it is assumed that the underlying value of the longitudinal outcome is associated with the survival outcome. However, in some cases there could be different characteristics of the patients’ longitudinal profiles that may influence the hazard, such as how fast the biomarker progresses. Therefore, not only the current value but also the slope of the longitudinal outcome could be included in the survival submodel. Furthermore, a summary of the whole history of the longitudinal outcome could be associated with the hazard. The choice of the functional form is an important task and need to be investigated since it could influence the results. In this work, we use a variable selection method under the Bayesian framework in order to obtain the most appropriate functional forms. Specifically, we propose a joint model that includes all the different association structures mentioned above and assume informative priors for the regression coefficients that correspond to the terms of the longitudinal process. We investigate different Bayesian priors, such as Bayesian lasso and ridge regression, that shrink small effects to zero. This is applied on a dataset consisting of patients with a chronic liver disease (primary biliary cirrhosis), where it is important to investigate which characteristics of serum bilirubin (presence of blood vessel malformations in the skin) have an influence on survival.
Joint modelling of longitudinal and multi-state processes: an application in patients with localized prostate cancer

L. Ferrer, James Dignam, Virginie Rondeau, Tom Pickles and Cecile Proust-Lima

Joint modelling of longitudinal and survival data is increasingly used in clinical trials on cancer. In prostate cancer for example, these models permit to account for the link between longitudinal measures of prostate-specific antigen (PSA) and the time of clinical recurrence when studying the risk of relapse. In practice, there are multiple types of relapse and some may follow. Distinguish these transitions between health states would allow to evaluate, for example, which covariates impact the risk of dying after a distant recurrence post-radiotherapy, or to predict the risk of one specific type of clinical recurrence post-radiotherapy, from the PSA history. In this context, we present a joint model for a longitudinal process and a multi-state process which is divided into two sub-models: a linear mixed sub-model for longitudinal data, and a multi-state sub-model with proportional hazards for transition times, both linked by shared random effects. Parameters of this joint multi-state model are estimated within the maximum likelihood framework using an EM algorithm coupled to a quasi-Newton algorithm in case of slow convergence. It is implemented under R, by combining and extending the mstate and JM packages. The multi-state data are prepared through the mstate package, and a slightly modified jointModel() function carries out the estimation procedure. The estimation program is applied on pooled data from two cohorts of men with localized prostate cancer and treated by radiotherapy. Thanks to the classical covariates available at baseline and the PSA measurements collected repeatedly during the follow-up, we are able to assess the trajectory of the biomarker, define the risk of transition between health states, and quantify the impact of the PSA dynamics on each transition intensity.

L. Ferrer
Inserm U897, Bordeaux, France
e-mail: loic.ferrer@isped.fr
Joint latent class model for longitudinal data and interval-censored semi-competing events: Application to Alzheimer’s disease

Anais Rouanet, Pierre Joly, Jean-Francois Dartigues, Cecile Proust-Lima and Helene Jacqmin-Gadda

Joint models are used in ageing studies to investigate the association between longitudinal markers, such as cognitive tests, and a time-to-event, such as time-to-dementia. These models have been extended to the study of multiple markers and/or competing risks. The competing risk of death must be considered when studying time-to-dementia in the elderly because the risk of death is high in this population, and death and dementia have common risk factors including cognitive decline. Moreover, in cohort studies, time-to-dementia is interval-censored because dementia diagnosis is only assessed at the follow-up visits. Subjects can thus get demented and die between two visits without being diagnosed as demented. To avoid bias, the risk of dementia must be estimated with an illness-death model that enables to account for a possible unobserved transition to dementia. To study pre-dementia cognitive decline considering both the competing risk of death and interval-censoring, we develop a joint latent class model combining a (possibly multivariate) mixed model and an illness-death model. Parameters are estimated by maximum likelihood handling interval-censoring. The correlation between the marker and the times-to-events is captured by latent classes, corresponding to homogenous subgroups of the population with specific risks of death and dementia and specific profiles of cognitive decline. Two versions of the model, markovian and semi-markovian, are proposed. This model is compared to a joint latent class model for standard competing risks through a simulation study. The method is applied on a French prospective cohort study of cerebral and functional ageing to distinguish different profiles of cognitive declines associated with the risks of dementia and death. The comparison of markovian and semi-markovian models highlights that mortality of demented subjects depends more on age than on delay in dementia. This model enables to distinguish the so-called terminal decline before death (among non-demented) from decline due to dementia.

___________________
Anais Rouanet
ISPED, Inserm Center U897, Bordeaux, France
e-mail: anais.rouanet@isped.u-bordeaux2.fr
Contributed Session C9: Survival, prediction, joint models
Estimating the type-specific cumulative incidence of cervical precancer in a setting with competing risks and masked event types

Birgit Witte and Johannes Berkhof

Cervical cancer is caused by a persistent infection with high-risk human papillomavirus (hrHPV). Approximately 20-50% of hrHPV-positive women are infected with more than one hrHPV-type but most hrHPV type infections are transient. Current cervical cancer prevention programs are likely to change due to innovations such as HPV genotyping tests and new vaccines targeting multiple hrHPV genotypes. To predict impact of current and future prevention programs, estimates of hrHPV type-specific (pre)cancer (CIN2+) risks are required. The link between hrHPV genotypes and CIN2+ is masked in women with multiple HPV type infections. In principle, the responsible hrHPV type can be identified by laser capture micro-dissection (LCM) technology (van der Marel, 2014). However, this technology is costly and therefore, the current large screening cohorts yield data that can be described by a competing risks model with masked event types. In the early 1980s, Dinse (1982) and Miyakawa (1984) were the first to study these competing risks models. Since then, several papers have been written for a variety of assumptions and situations. These models, however, assume that every subject under study is at risk for each of the possible risks, and therefore do not address the most important issue that arises in the cervical cancer prevention setting: the relative CIN2+ risk for HPV types not detected at baseline is nearly zero. We derive the likelihood in an exponential right-censoring model that accounts for possibly varying risk sets amongst subjects. In a simulation study we show that the maximum likelihood estimates (MLEs) of the cause-specific risks, obtained with Newton’s algorithm, are consistent. Moreover, we compare estimates from our model in an example with 257 CIN2/3 cases where the missing link between the competing risks and event type is recovered via LCM technology. We end with directions for future extensions, including interval censoring methods.

Semiparametric transformation models (STM) form a flexible class of regression models to analyse time-to-event data. Special cases include the Cox proportional hazards model and the proportional odds model. The general form of the STM lacks an informative effect measure. Here we propose to measure covariate effects on the probabilistic index scale. The probabilistic index quantifies the probability that the outcome of one subject exceeds that of another conditional on covariates. We show that this effect measure can be derived under any element of the class of STMs. We then discuss model based estimation as well as estimation when the model is misspecified. Furthermore we characterise the relation between the STMs and the class of probabilistic index models.
Multivariate mixtures of Erlangs for density estimation under censoring and truncation

Roel Verbelen

We study the estimation and use of multivariate mixtures of Erlang distributions (MME) to model dependent multivariate censored and truncated data. MME form a highly flexible class of distributions as they are dense in the space of positive continuous multivariate distributions. Moreover, the class is analytically tractable. Many quantities of interest such as the joint density and distribution function, the Laplace transform, moments, Kendall’s tau and Spearman’s rho have a closed form. The class also enjoys appealing closure properties such as the facts that any uni- or multivariate marginal or conditional distribution is a uni- or multivariate Erlang mixture, the distribution of the sum of the component random variables is a univariate Erlang mixture and the distribution of the residual lifetimes is a again a multivariate mixtures of Erlangs. The use of MME should be regarded as semiparametric density estimation technique to model the dependence directly and hence forms a suitable alternative to the use of copulas. We present an estimation technique for fitting MME using the EM algorithm to data that can be censored and/or truncated, which is often the case in applications such as clinical experiments (survival / failure time analysis), mastitis studies (veterinary studies), loss modeling (finance and actuarial science) and duration data (econometric studies). We demonstrate the effectiveness of the proposed algorithm and the practical use of MME on simulated data as well as on real-life data from a mastitis study.

Roel Verbelen
Katholieke Universiteit, Leuven, Belgium
e-mail: roel.verbelen@kuleuven.be
Event-based Ascertainment of Recurrent Events Data

Theodor Adrian Balan, M.A. Jonker, and H. Putter

Recurrent events are increasingly common in clinical studies. Often, such data is collected retrospectively from hospital registries. In this case, the ascertainment of the patients can depend on the history of the recurrent event process; for example, only patients who have experienced at least one event during a pre-specified (calendar) time period are selected. We propose a flexible model and a likelihood-based procedure which takes the selection process into account when estimating the model parameters. In the context of recurrent events, it is known that the selection conditions must be considered in developing an analysis (Cook and Lawless 2007), however a general approach appears to be lacking. Furthermore, the event-based ascertainment criteria seem to be ignored in several studies of recurrent risks, and although it has been suggested that this might be the source of some paradoxes in clinical research (Dahabreh and Kent 2011), the consequences on the estimation procedure have not been extensively studied. The same holds in particular for frailty models, in which a random effect is used to explain unaccounted heterogeneity. We adapt existing frailty models for recurrent events by accounting for the event-based ascertainment, while analysing the pitfalls associated with ignoring the selection process. We derive consistent estimators of the model parameters under different ascertainment schemes and study the small sample properties of these estimators through a simulation study. Finally, we illustrate the proposed methods on a data set comprising of recurrent pneumothoraces, for subjects who were ascertained if they had experienced at least one event during a certain time window.
Dynamic frailty models based on compound birth-death processes

Hein Putter and Hans van Houwelingen

Frailty models are used in survival analysis to model unobserved heterogeneity. They accommodate such heterogeneity by the inclusion of a random term, the frailty, which is assumed to multiply the hazard of a subject (individual frailty) or the hazards of all subjects in a cluster (shared frailty). Typically, the frailty term is assumed to be constant over time. This is a restrictive assumption and extensions to allow for time-varying or dynamic frailties are of interest. In this presentation we extend the auto-correlated frailty models of Henderson and Shimakura and of Fiocco, Putter and van Houwelingen, developed for longitudinal count data and discrete survival data, to continuous survival data. We present a rigorous construction of the frailty processes in continuous time based on compound birth-death processes. Emphasis will be on the construction of the processes. Properties of survival models where the frailty processes are used as mixtures will briefly be discussed.

Hein Putter
Leiden University Medical Center, Leiden, The Netherlands
e-mail: h.putter@lumc.nl
Direct modeling of regression effects for transition probabilities in the non-Markov illness-death model

Leyla Azarang, Thomas Scheike, and Jacobo de Uña-Alvarez-Alvarez

Multi-state models are often used to represent the individuals’ progress along a certain disease. The estimation of transition probabilities is an important goal in such a setting. The progressive illness-death model is an important multi-state model which has many applications in medical research. In this work, we introduce direct binomial regression to model the covariate effects on transition probabilities in the (possibly non-Markov) censored progressive illness-death model. To this end, a binomial regression model with time-varying effects is considered, and an estimator of the time-varying coefficient is introduced. The performance of the proposed estimator is investigated through simulations. Also, for illustration purposes, a real medical dataset is analysed.

Leyla Azarang
University of Vigo, Vigo, Spain & University of Vigo, Vigo, Spain
e-mail: leyla.azarang@uvigo.es
Contributed Session C10: Ecology
Abundance estimation using integrated modelling of unmarked animal data from camera traps

Natoya Jourdain, Diana J. Cole, Martin S. Ridout, and Marcus J. Rowcliffe

The use of camera traps to estimate animal density has so far been restricted to capture-recapture analysis of species with unique identifiable markings. The Random Encounter Model (REM) developed by Rowcliffe et al., (2008) estimates animal abundance without the need for individual recognition of the animals. The REM describes rates of contacts between animals and camera traps from which an estimator for animal density can be derived. To obtain an estimate of population density, the REM considers two distinct and independent sets of data: 1) trap rate data, and 2) an estimate of animal speed of movement, which is a fixed average. We extend the REM by developing an integrated likelihood. The extended Random Encounter Model (eREM) adopts maximum likelihood estimation, modelling simultaneously camera trap and animal speed of movement data. We modelled the camera trap data by a Poisson distribution and animal speed by a gamma distribution. Testing the reliability of this approach through simulation, we found that eREM is an improvement over REM. Our analyses confirm that the REM provides approximately unbiased estimates of the parameters themselves but underestimates their variances, whereas the eREM is unbiased and estimates variance correctly. The eREM was extended to include covariates such as habitat type and camera. In the analysis of a real data set, the study site was divided into four areas of contrasting habitat. The results showed that density varied with habitat. Also, what is interesting is the estimation of non-zero densities for some species in habitats where there were no census observations.

Estimation of $P(X > Y)$ for normal distributions in the context of probabilistic environmental risk assessment for nanoparticles

Rianne Jacobs, Andriette Bekker, and Cajo ter Braak

As is the case for all novel materials, risk assessment (RA) is important for the societal acceptance and safe use of engineered nanoparticles. In order to perform a proper RA, we need knowledge and data on the properties of nanoparticles. This information is hard to come by due to lack of knowledge and technical limitations, resulting in large amounts of uncertainty. When high levels of uncertainty are foreseen, a probabilistic RA is recommended. In probabilistic RA, variability in environmental exposure is quantified by an exposure distribution. Similarly, variability in effect is quantified by a species sensitivity distribution (SSD). The overlap of the exposure distribution and the SSD forms the basis for risk characterization in a probabilistic RA and is defined as the risk probability, $R = P(X > Y)$. In this paper, we consider three parametric estimators of the risk probability. These are the maximum likelihood estimator (MLE), quasi maximum likelihood estimator (QMLE) and Bayesian estimator with a diffuse prior. We also provide the non-parametric empirical estimator as comparison. Monte Carlo simulation was performed for combinations of sample sizes and R values to evaluate the performance of the estimators. We conclude that for small sample sizes the MLE and QMLE give the best results. The non-parametric estimator is incapable of estimating small R values. Even for the largest sample sizes that we used the estimator cannot estimate R values lower than 0.0001. The Bayesian estimator, although very biased for lower sample sizes, has a high accuracy compared to the other estimators for small R values. This could possibly be interesting to explore when a more informative prior is used, based on existing prior knowledge.
Modelling the sea lice population at a salmon farm

Magne Aldrin

Parasitic salmon lice are potentially harmful to salmon hosts and farm produced lice pose a threat to wild salmon and sea trout. To control salmon lice infections in Norwegian salmon farming, numbers of lice are counted and reported from all salmon farms every week. If the lice abundance at a farm becomes too high, e.g. more than 0.5 adult female lice per fish, medical treatments are used to control the lice abundance. In addition non-medical treatments are used, the most important being cleaner fish. We have developed a population model for lice at a salmon farm, formulated as a Bayesian hierarchical model. In this model, the life cycle of lice is divided into five stages; i) Recruits (eggs and small larvae drifting in the water), ii) copepodits (larvae that are able to find a fish host), iii) chalimus (sessile lice on a fish), iv) pre-adults (mobile lice on a fish) and v) adults (also mobile lice on a fish). The adults are further divided into males and females. A louse in a given stage will from one day to the next either a) remain in the same stage, or b) go over to the next stage, or c) die or drift away from the farm. Furthermore, adult females generate new eggs. In addition to self-generated recruitment there will be externally infestations of lice due to infection from neighboring farms. The development rates from one stage to another depend on the sea temperature, whereas the mortality comprises both natural mortality and mortality due to treatments. The parameters in the model are estimated using data from about 20 fish farms and an MCMC algorithm. We will provide selected results from the model and demonstrate how future lice abundance can be predicted conditioned on given scenarios for future treatment against lice.
To be able to fit or examine a parametric model successfully all the parameters need to be identifiable. If the parameters are non-identifiable the model can be rewritten in terms of a smaller set of parameters, and is termed parameter redundant. Parameter redundancy is not always obvious, in which case the definitive method for detecting parameter redundancy is a symbolic method. This involves calculating the rank of a matrix, which is expressed symbolically (see for example Cole et al., 2010, Mathematical Biosciences, 228, 16-30). Many studies of ecological populations consist of more than one type of data. Such data sets can be examined simultaneously using an integrated population model. However combining several data sets does not necessarily mean all the parameters can be estimated. We present methods for determining which parameters can be estimated in integrated population models. Models used in ecology are becoming more realistic but at the same time more complex. This poses a problem for the symbolic method as computers run out of memory calculating the rank of the appropriate matrix. Combining more than one model in an integrated population model adds to this complexity. We present extensions to the symbolic method that simplify the calculation considerably. The methods are illustrated using an integrated population model for four data sets on Common Guillemots. This involves capture-recapture data, mark-resight-recovery data, productivity data and count data.
Contributed Session C11: Survival, prediction, joint models
Dynamic prediction by direct binomial regression

Mia Klinten Grand and Hein Putter

In recent years there have been a series of advances in the field of dynamic prediction: Among those is the development of methods for dynamic prediction of the cumulative incidence function in a competing risk setting. These models enable the predictions to be updated as time progresses and more information becomes available, e.g. when a patient comes back for a follow-up visit after completing a year of treatment, the risk of death and adverse events may have changed since treatment initiation. We extend the direct binomial regression approach by Scheike et. al (2008) to model the dynamic cumulative incidence function and show the asymptotic properties of the estimators. The proposed models are very flexible as they allow the covariates to have complex time-varying effects and we introduce tests to investigate these possible time-varying structures. The method handles incompletely observed event times by inverse probability weighing of individuals who have experienced the event of interest and the models can be fitted by generalized estimating equations. One advantage of the method, compared to modeling the cause-specific hazards, is the direct interpretation of the effect of the covariates on the cumulative incidence. Finally we illustrate the method on real data and compare it to the other existing methods based on cause-specific hazards and dynamic pseudo-observations.

Mia Klinten Grand
Leiden University Medical Center, Leiden, the Netherlands
e-mail: M.Klinten_Grand@lumc.nl
Dynamic prediction of mortality amongst patients in intensive care using the sequential organ failure assessment (SOFA) score

Jammbe Z Musoro

In intensive care units (ICU), besides routinely collected admission data, a daily monitoring of organ dysfunction using scoring systems such as the Sequential Organ Failure Assessment (SOFA) score has become practice. Such updated information is valuable in making accurate predictions of patients' survival. Few prediction models that incorporate this updated information have been reported.

Method: We looked at follow-up data of ICU patients who either died or were discharged at the end of hospital stay, with no censored cases. We propose a joint model comprising a linear mixed effects submodel for the development of longitudinal SOFA scores, and a proportional subdistribution hazards submodel for death as end point with discharge as competing risk. The two parts are linked by shared latent terms. Since there is censoring, it is easy to calculate performance measures such as the Brier score and the AUC. We compared predictive values from our model with those obtained from an earlier modelling approach by Toma et al. (2007).

Results: It was straightforward to fit and calculate survival probabilities in the joint modelling framework. No a priori computationally demanding search for the best predictors, as with the Toma approach, was needed. Similar predictive performances were obtained from both approaches, though the joint model sometimes outperformed the Toma approach.

Conclusion: The joint modeling framework can serve as an additional predictive tool to make decisions as regards to caring of ICU patients.
Time dependent discriminative ability of dynamic prediction models with discrete survival times

N. van Geloven and A.H. Zwinderman

Dynamic prediction models which incorporate longitudinal covariate data and allow for temporal updating of predictions are increasingly popular. Such models allow repeated predictions over time using all information available up to that time point. Several methodologies have been developed for dynamic prediction, including landmark approaches and joint models. As a result of the entry of these analysis techniques, attention has been given to the evaluation of the performance of dynamic prediction models. The incident/dynamic time-dependent concordance index proposed by Heagerty and Zheng (2005) seems an appealing measure to express the discriminative value of dynamic survival models. Estimation of this measure was proposed using three different model assumptions for the relation between the dynamic (linear) predictor and the event. Van Houwelingen and Putter (2011) propose a model-free alternative of this measure, focussing on applications where the survival times can be measured in continuous time, precluding tied event times. We show how this model-free concordance measure can be calculated in case of discrete event times, allowing for ties. We illustrate the application of the proposed measures using data form a large (n<5000) cohort study focussing on the discrete outcome number of menstrual cycles to pregnancy. Using this case study we show the numerical and interpretational differences and similarities of the measures.

Nan van Geloven
Academic Medical Centre, Amsterdam, the Netherlands
e-mail:n.vangeloven@amc.uva.nl
Contributed Session C12: Longitudinal data & Causality
Multiple imputation for incomplete predictors and complex outcomes: mice vs. sequential Bayesian imputation

Nicole Erler, Dimitris Rizopoulos, Oscar H. Franco, and Emmanuel M.E.H. Lesaffre

This work is motivated by a large cohort study that investigates the effect of sugar-sweetened beverages on longitudinal BMI in toddlers. A major challenge in the analysis of this dataset is that many predictive factors have missing values (39% complete cases). An effective procedure to overcome the incomplete covariate information problem is multiple imputation (Rubin, 1987). Within this context, the approach of Multiple Imputation with Chained Equations (MICE, Van Buuren, 2012) has been shown to be quite successful and effective in creating complete datasets that can then be analyzed with the preferred analysis method. The basic idea behind MICE is to define a series of regression models that link each incomplete predictor with other predictors and with the outcome. When defining these models for our data, it was not directly evident how the longitudinal outcome should be incorporated into these predictive models. Longitudinal outcomes often also have missing data, and not all subjects provide measurements at the same time points. Different specifications resulted in different conclusions, indicating that it can be crucial how a complex outcome is included in these equations. To avoid the need to specify appropriate predictive models for all incomplete predictors we propose a full Bayesian approach that combines the imputation and analysis parts. In this framework we directly obtain inferences for the full posterior distribution of the parameters and missing covariates, with the likelihood of the model decomposed in a series of univariate conditional distributions (Ibrahim et al., 2002). Another important advantage over MICE is that Bayesian imputation provides valid inferences under MAR (missing at random) missing data mechanisms. Extensive simulation comparisons of MICE versus the full Bayesian approach corroborated the fact that for the former analysts need to correctly specify the longitudinal process in the predictive models, whereas the latter performs well under different scenarios.

Nicole Erler
Erasmus University Medical Centre, Rotterdam, Netherlands
e-mail: n.erler@erasmusmc.nl
Pseudo-Likelihood Approach for Large and Complex Ordinal Data

Anna Ivanova, Geert Molenberghs, and Geert Verbeke

In longitudinal studies, continuous, binary, categorical, and survival outcomes are often jointly collected. However, when it comes to modelling responses, the ordinal ones have received less attention in the literature. In a longitudinal or hierarchical context, the proportional odds mixed model (POMM, a model for ordinal responses) can be regarded as an instance of the generalized linear mixed model. When the responses of the joint multivariate model encompass ordinal responses, the complexity further increases. An additional problem of model fitting is the size of collected data. Pseudo-likelihood based methods for pairwise fitting (Fieuws et al., 2006), for partitioned samples (Molenberghs et al., 2011) and, as introduced in our work, a combination thereof, enabled us to jointly model large amounts of responses. To illustrate our methodology, we used the composite endpoint on an ordinal scale taken from a diabetes study: multiple targets for HbA1C, LDL-cholesterol and SBP. For every response, a univariate POMM was formulated. To capture the association between the separate responses, an assumption about the correlation between random effects was made. The following observations were made. First, alternative pseudo-likelihood methods yield valid estimates with high efficiency, even for low numbers of quadrature points (Q = 3). The big advantage of these methods is their gain in computation time over the full likelihood method. We achieved a significant reduction in computation time: from 7 minutes 13 seconds (for full likelihood method) to only 20 seconds (for the combined method). This was because the submodels could run in parallel.


Anna Ivanova
Katholieke Universiteit Leuven, Leuven, Belgium
e-mail: anna.ivanova@lstat.kuleuven.be
Prediction of glaucoma progression using a high-dimensional two-stage Bayesian hierarchical model


Our research is motivated by an ongoing Dutch study, conducted by the Rotterdam Eye Hospital, of the longitudinal series of visual fields (VF) in glaucoma patients. The evaluation of these VFs may allow early detection of glaucoma and functional deterioration. Two challenging characteristics of the VF measurements collected in this study are that they are left-censored and that they have a four level hierarchical longitudinal structure, namely, for each individual (level 1) we have separate measurements per eye (level 2), per hemifield within each eye (level 3), as well as for each of the 26 locations within each hemifield (level 4). Our initial attempts to fit the full hierarchical Bayesian model were not successful due to the high dimensionality of the data that resulted in long computing times and computer memory problems. To overcome these difficulties and to still be able to fit the model in a realistic timeframe, we have developed an extended version of the recently proposed two-stage approach of Lunn et al. (2013). More specifically, we split the model at the individual level and in Stage 1 analyze each individual separately. Following in Stage 2, we use the sampled parameters from the previous stage as proposal distributions to obtain population-averaged estimates. Thus, we are able to remove the computational burden while still benefiting from a full Bayesian model. One disadvantage of this approach is that it does not provide the required components to evaluate the fit and predictive ability of the model using the Deviance Information Criterion (DIC). In order to calculate the DIC and compare different competing models for our data fitted using the two-stage approach, we will illustrate a Monte Carlo scheme based on a Metropolis within Gibbs algorithm.

Susan Bryan
Erasmus Medical Center, Rotterdam, Netherlands
e-mail: s.bryan@erasmusmc.nl
Contributed Session C13: Ecology
Traditionally, the modelling of the spatial effects in field trials has used correlated random noise. To make this work, simplifying assumptions, like a separable autoregressive structure, have to be made. In practice a lot of tuning is needed to get results. We propose to model the spatial field with tensor products of B-splines. In the spirit of P-splines we use anisotropic difference penalties to tune smoothness. A special reformulation leads to six additive spatial components. They form the basis of a mixed model with six unknown variance parameters. On top of the field come fixed or random effects for genotypes, rows and columns and possibly sub-regions of the field. Although the model contains up to ten variance components, estimation is stable and fast, using an EM-like algorithm. Hundreds of fields, with thousands of plots each, have been analyzed without human intervention. A package for R is available. Our experience has shown that the effective dimensions connected to the model components play a crucial role in the estimation process. They are also very useful to summarize results. We illustrate our work with challenging experimental data.
Latent Block Model for ecological abundance data

Julie Aubert, S. Schbath and S. Robin

Binary or count matrices are widely used in numerous fields and namely in ecology. Metagenomics which studies microbial communities directly from environmental samples, provides abundance matrices where rows correspond to bacteria and columns to sampling units (SU). One major goal is to find associations between bacteria communities and SU. We propose a model and associated inference procedures for a simultaneous clustering: one on the populations of bacteria constituting the metagenome, and the other one on the SU. To this aim we use the latent block model framework introduced by Govaert and Nadif (2010). We assume that the two latent variables corresponding to the groups are independent and follow multinomial distributions. We assume that conditionally on the unknown labels of groups respectively in rows and columns the abundance data are independent and follow a certain distribution. We parametrize this distribution in order to take into account the sampling effort and the average species abundance across all units. The aim of the inference is to estimate both the hidden variables and the parameters. Since latent variables are not independent conditionally on observed variables, the classical maximum likelihood inference is intractable. Govaert and Nadif (2010), Keribin et al. (2014) suggest as alternative inference algorithms based on a variational approach (Wainwright and Jordan (2008)). This approach consists in maximizing a lower bound of the log-likelihood of the observed data using a tractable distribution. Recent studies show that metagenomics data are overdispersed. We thus extend a Poisson model more natural when dealing with abundance matrices to a Poisson-Gamma model (which induces a third hidden layer) and propose generalized Variational Expectation-Maximization algorithms for inference in both cases. We apply them to a rhizosphere metagenomics dataset.
Ultrasonic vocalizing behaviour in socially housed rats: a composite-link function approach

Nadia Vendrig, Ilona Pinter, Suzanne Peters, Berry Spruijt and Cajo ter Braak

Studying laboratory rats is one of many steps towards understanding and potentially curing neurological and psychiatric diseases in humans. If a symptom of such a disease is social impairment we wish to study rats in a social context. In this study, we aim to relate high frequency ultrasonic vocalisations (USVs) to (social) behaviour. Rats were housed in pairs in an automated home cage system (PhenoTyper, Noldus) which records the exact location of both rats continuously. From these data, various behaviour parameters were calculated per individual rat. However the cage was equipped with a single microphone, so that USVs could not be assigned to individual rats. This created a statistical challenge of how to relate USVs to behaviour from such data. We met the challenge using a composite-link function approach which models the number (or rate) of uttered USVs per pair of rats as the sum of two expected values, each linked via a log (or complementary log-log) link function with its own linear predictor. This differs from the traditional GLM approach where the total number of uttered USVs per pair is predicted by one linear predictor that includes behaviour parameters of both rats. From a biological point of view, the composite link-function approach is more natural, because USVs are uttered by individual rats rather than pairs of rats. Our simulation study indicates that the composite link-function outperforms traditional GLM in this setting. Furthermore, the composite link-function approach can accommodate more flexible parametrisations to test specific hypothesis that would not be straightforward in a log-link setting.

Nadia Vendrig
Biometris, Wageningen UR, Wageningen, Nederland
e-mail: nadia.vendrig@wur.nl
Capture-recapture and the penalized composite link model

Paul Eilers

One type of capture-recapture problems leads to discrete distributions with a missing frequency at zero, because it is only possible to count cases that occur at least once. An estimate of the missing frequency can be obtained if we assume that we are dealing with a truncated parametric distribution. Poisson and negative-binomial are popular choices, but their fit to the data might not be good. I propose to use a mixture of Poisson distributions, with a smooth non-parametric mixing distribution. It is a special case of the penalized composite link model, leading to an elegant and efficient estimation procedure. To illustrate the model I use scrapie data from the UK, as reported in a paper by Boehning et al., who use finite mixtures. Scrapie is a sheep disease; cases have to be notified to the authorities. It is important to have an estimate of the amount of underreporting. The data are available for many counties. I will explore opportunities to borrow strength from neighbors to get more stable estimates for counties with low numbers of cases.
Using pleiotropic genetic variants to simultaneously estimate multiple causal effects: a multivariable extension to Mendelian randomization

Simon G. Thompson and Stephen Burgess

A conventional Mendelian randomization analysis assesses the causal effect of a risk factor on an outcome by using genetic variants solely associated with the risk factor of interest as instrumental variables. However, in some cases it may be difficult to find a relevant genetic variant that is not also associated with related risk factors. Such a variant is known as pleiotropic. We describe an extension to Mendelian randomization that uses multiple genetic variants associated with several measured risk factors to simultaneously estimate the causal effects of each of the risk factors on the outcome, and discuss the assumptions necessary for a valid analysis. This multivariable Mendelian randomization approach is similar to the simultaneous assessment of several treatments in a factorial randomized trial. We present a two-stage least squares method for individual-level data, a likelihood-based method for summarized data, and a weighted regression-based method for summarized data. We show that these produce almost identical results in simulations. The methods are applied to disentangle the causal effects of low-density lipoprotein cholesterol, high-density lipoprotein cholesterol and triglycerides on coronary heart disease.
One of the basic aims in science is to unravel the chain of cause and effect of particular systems. Especially for large systems this can be a daunting task. Detailed interventional and randomized data sampling approaches can be used to resolve the causality question, but for many systems such datasets are impossible or too costly to obtain. Recently, work by Maathuis et al (Nature Methods, 2010), following ideas from Spirtes and Glymour, has introduced a framework to estimate causal effects in large-scale Gaussian systems. By describing the causal network as a directed acyclic graph it is possible to estimate a class of Markov equivalent systems that describe the underlying causal interactions consistently. Our work extends this idea to general, non-Gaussian systems. In these systems, causal effect stop being linear and need to be estimated functionally. By means of a Copula approach, the underlying causal functional can be written sparsely and estimated effectively. This method is applied to a relatively large, observational gene expression dataset.
The dynamic approach to causal reasoning in ageing studies

Daniel Commenges

We consider the issue of estimating causal effects in a dynamic approach based on a multivariate stochastic process representation, which may be called the stochastic system approach. Conditional and marginal effects can be defined. We focus on the issue of the horizon on which causal influence must be studied, in particular in ageing studies. In ageing studies, one of the most important events that we have to consider is death. This is why the illness-death model is important in such studies. But death is not an event which is on the same footing as other events that can happen to subjects. Even if the vital status is part of the state, it has a very special meaning, in that all the other components of the state are defined only for a living subject. The consequence is that causal influences must be defined on a maximum horizon which is the time of death. We do not say that death has an influence on the other components of the state, but that these other components are not defined after death. For instance if we are interested in dementia, the state can be represented by a bivariate counting process counting dementia and death. However dementia is defined only for a living subject: after death the subject does not exist anymore and cannot be qualified as demented or not demented. When we investigate the causal influence of a factor, we should first look at its causal influence on death, then on its influences on other processes. Cases where a value of a modifiable factor can be preferred to another one will be given. Thus, the stochastic system approach helps clarifying the important issue of assessing causal effects in ageing studies.

Daniel Commenges
INSERM, Bordeaux, France
e-mail: daniel.commenges@isped.u-bordeaux2.fr
Fitting linear mixed models with and without centering the explanatory variables

Alan Welsh and Hwan-Jin Yoon

Linear mixed models are widely used to describe the relationship between a response and explanatory variables when the data are clustered. The effect of the cluster structure in the response variable is incorporated by including random effects in the model; cluster structure in the explanatory variables is often ignored, possibly because interpreting the model conditionally makes it seem unnecessary to consider this structure. At least when the explanatory variable has a normal distribution, cluster structure in the explanatory variable can be incorporated into the linear mixed model by centering the explanatory variable about the cluster means and then also including the cluster means in the model as a cluster-level covariate. We consider in detail the simplest two-level case (i.e. with only a single level of clustering) with a single explanatory variable and explore the effect of within-cluster correlation in the explanatory variable on estimates of both the regression parameters and the variance components obtained by maximum likelihood, reduced or restricted likelihood (REML) and h-likelihood estimation. We report a number of unexpected findings.(i) Not centering the explanatory variable when we should affects the estimates of both the regression and variance parameters not just the estimates of the regression parameters.(ii) Changing the within cluster correlation of an uncentered explanatory variable (by increasing the between cluster standard deviation of the explanatory variable) has different effects on different kinds of estimators. The least squares and weighted least squares estimators of the regression parameters are known to be smooth functions of the between cluster standard deviation of the explanatory variable but we show that the maximum likelihood and REML estimators of the regression and variance parameters are discontinuous functions of this standard deviation.(iii) The jump discontinuity in the maximum likelihood and REML estimators when fitting the linear mixed model with an uncentered explanatory variable occurs at different values of the between cluster standard deviation of the explanatory variable.(iv) Standard statistical software such as SAS, SPSS, STATA, lmer and GenStat often returns incorrect maximum likelihood and REML estimates in this very simple problem because they find a local maximum rather than the global maximum.

Alan Welsh
Australian National University, Canberra, Australia
e-mail: Alan.Welsh@anu.edu.au
Variable selection and deselection in regression models, with confidence

Aldo Solari and Jelle J. Goeman

It is common practice in regression models to perform data-driven variable selection and to draw inference from the resulting model as if it was given a priori (Berk et al. 2013). However, testing the model on data that gave it birth is almost certain to overestimate its performance, for the optimizing process that chose it from many possible (Mosteller and Tukey, 1977). Indeed, p-values and confidence intervals from the selected model do not enjoy the inferential validity that is guaranteed when the model has been chosen a priori. The reason for the invalidity is that the selection process produces a model that is random in itself. If a valid evaluation of a selected model is the objective, the uncertainty that comes along with the selection from many possible models should be accounted for. Post-selection inference is therefore essential for variable selection and deselection with confidence. We propose confidence statements for number of truly relevant covariates present in a selected subset, using the closed testing procedure (Goeman and Solari, 2011). The confidence statements we obtain are simultaneous for all possible subsets, so that they are not affected by the selection that is inherent in variable selection procedures.

Aldo Solari
University Milano-Bicocca, Milano, Italy
e-mail: aldo.solari@unimib.it
Exact testing with random permutations

Jesse Hemerik and Jelle J. Goeman

Permutation tests are hypothesis tests than can be used when a group of transformations exists that leave the distribution of the data invariant under the null hypothesis. It is necessary that the set of transformations is a group, although there are some exceptions. Often the number of transformations is too large, leading researchers to use a random subcollection of the group. This is based on the intuition that some subsets will give a too large rejection probability under the null hypothesis and other ones a too small one, but on average they give a rejection probability of . This is only approximately true however: the identity should always be added to the collection of random transformations. This is only very rarely stated (Ge, Dudoit, Speed (2013), p.16). Not adding the identity transformation can lead to somewhat too small p-values, which can give serious problems in multiple testing contexts. An example is given by Phipson and Smyth (2010). They have given correct p-values for both the case of drawing random permutations without and the case of drawing with replacement. In the former case, from their formula it can be seen that the test becomes exact if one adds the identity. We have shown that a similar result holds for the case of drawing with replacement. There are various multiple testing methods that are based on the permutation test. Examples are Westfall and Youngs maxT method using random permutations and Meinshausens (2006) procedure for finding a lower bound (uniform over the rejected sets) for the amount of correct rejections. A third example is a permutation-based global test that we introduce. All these methods become anti-conservative if we do not add the identity. Especially the new global test then becomes very anti-conservative, even if we use many random transformations. We have proven that the identity should be added in all these methods. We have applied these procedures to simulated data to illustrate the importance of adding the identity.

Jesse Hemerik
Radboud university medical center, Nijmegen, The Netherlands
e-mail: Jesse.Hemerik@radboudumc.nl
Count data analysis in nutrition clinical trials

Yalcin Yavuz and Sophie Swinkels

In Early Life Nutrition division of Danone Nutricia Research, the analysis of count data such as the number of infections, number of hospital visits, number of doctor-diagnosed diarrhea, is of interest in many clinical trials. Poisson regression models would be the simplest standard framework for the analysis of such data. However, in real life, count data do not always meet the assumption of equal variance-mean relationship induced from Poisson distribution leading to over-(or under-) dispersion. The source of the over- (or under-) dispersion could be due to a higher than expected occurrence of zero counts. A toddler may have no infection either because of his/her resistance to the infection, or simply because no disease spores have landed on him/her. This is the distinction between structural zeros, which are inevitable, and sampling zeros, which occur by chance. Another source of over-dispersion might be the fact that having an infection might make individuals more vulnerable for a second one. We demonstrate the use of four different models for over-dispersed count data with certain levels of zero inflation: Poisson, negative binomial, zero-inflated Poisson and zero-inflated negative binomial models. We discuss the performance of the models using data from a clinical trial on the number of infections in toddlers during a 12 month period.

Yalcin Yavuz
Danone Nutricia Research, Utrecht, The Netherland
e-mail: yalcin.yavuz@danone.com
Poster lightning presentations P1
Best-after-breast design: challenges of nutrition intervention studies in infants

Elleny Balder and Sophie Swinkels

Breastfeeding is the preferred and recommended method of infant feeding. If a mother cannot or chooses not to (fully) breastfeed, a commercially prepared infant formula is the recommended alternative. This presentation will address the challenges of nutrition intervention studies in infants, due to practical, ethical and regulatory issues that are inherent to studying feeding regimens in infants. A Best-after-breast design was implemented in Danone Nutricia Research, which allows subjects to enter the study without interfering with the mother’s or parents’ choice of early nutrition for the infant. After inclusion (28 days of age), regardless of the feeding regimen, subject data is recorded. When the mother/parents autonomously decide to start formula feeding, the subject is randomized in a double-blind parallel design to one of the study products. After start with the study formula, the mother is free to continue breast feeding in combination with formula as long as she wants and/or she can switch to full formula feeding in her own pace at any time. The different feeding regimens in the Best-after-breast design may introduce time-varying confounding. To account for this, for each subject for each measurement the state and duration of full breast feeding, mixed feeding (combination of formula with breast milk and/or weaning foods) and full formula feeding is determined. However, the choice to breast or formula feed may introduce selection bias that cannot be eliminated in the statistical analysis. These and other challenges that are encountered in infant nutrition intervention studies and possible solutions will be presented.

Elleny Balder
Danone Nutricia Research, Utrecht, The Netherlands
e-mail: elleny.balder@danone.com
The self-controlled case series (SCCS) method is used to investigate potential associations between vaccine or drug exposures and adverse events. It requires information only on cases, individuals who have experienced the adverse event at least once, and automatically controls all fixed confounding variables. Time-varying confounders such as age and season are not automatically controlled and must be modelled. The SCCS method has been extended by modelling only the age effect or only the time-varying exposure effect using splines while representing the other by a piecewise constant step function, so there is a need to pre-specify exposure groups or age groups a priori. Here, we present a non-parametric SCCS method in which both age and exposure effects are represented as linear combinations of cubic M-splines at the same time. Simulation studies showed that the new method performs well.
Modelling the effect of vaccinating boys along with girls against hpv

Johannes Berkhof and Johannes Bogaards

In many European countries, 12-year old girls are currently vaccinated against HPV 16 and 18 to protect women against cervical cancer. Inclusion of boys into the HPV vaccination programme is expected to provide some benefit to women, although the benefit will be small if programmes already achieve high coverage among girls. Protection of women should however not be the sole public health objective of HPV vaccination, as men are at risk of vaccine-preventable cancers themselves. Vaccination of boys is expected to lead to a reduction in vaccine-preventable cancer burden in men but the impact will strongly depend on the indirect benefit received from female vaccination. We will present a Bayesian evidence synthesis approach to evaluate the impact of vaccination against HPV types 16 and 18 on the burden of anal, penile, and oropharyngeal cancers among heterosexual males and men who have sex with men (MSM). A key component of the analysis is a HPV transmission model estimated from Dutch HPV type prevalence, cancer registry, and sexual activity data.

Johannes Berkhof
VU University Medical Center, Amsterdam, The Netherlands
e-mail: h.berkhof@vumc.nl
Individual patient data meta-analysis for diagnostic test accuracy studies: a review of methods used in practice

Junfeng Wang

Individual Patient Data (IPD) meta-analyses are regarded as the gold standard for systematic reviews. This conclusion also applies to systematic reviews of diagnostic test accuracy (DTA) studies. In recent years, an increasing number of DTA systematic reviews with IPD meta-analysis have been done. But so far there are no standard methods for data collection and statistical analysis in DTA IPD meta-analyses.

Methods: A systematic search was performed to identify published articles containing IPD meta-analyses of DTA studies. Embase and Medline databases were searched from 2000 to 2014. Data was extracted with a data extraction form developed for this study including both multiple choices and open questions. Results: 29 DTA IPD meta-analyses articles published between 2006 and 2014 were selected as the subjects for the final analysis. More than 80% (24/29) of the DTA IPD meta-analyses were performed based on datasets containing individual patient information provided by authors of primary studies. 60% (15/25) of the IPD meta-analyses included data from less than 10 primary studies. 60% (15/25) of the IPD meta-analyses had fewer than 2000 patients and only 1 meta-analysis included more than 10000 patients. Sensitivity and specificity and AUROC are most commonly used measures of test accuracy in DTA IPD meta-analysis. Regression models including fixed and random effects, multilevel and GEE logistic regression models are most commonly used statistical methods in DTA IPD meta-analysis. Additional analyses, which include subgroup analysis, covariate analysis and cut-off value analysis, added value to the DTA IPD meta-analysis.

Conclusion: Although the number of IPD meta-analyses in DTA systematic reviews is increasing, there is much variation in how these IPD MA were performed and what statistical methods were used. In this study, we showed the wide variation in methods used.

Junfeng Wang
Academic Medical Center, Amsterdam, Netherlands
e-mail: wangjunfeng7@gmail.com
Bayesian Geoadditive Regression and Latent Variable Models of Childhood Malnutrition in Gamo Gofa Zone

Tilahun Ferede Asena and Derbachew Asfaw

Major progress has been made over the last decades in reducing the prevalence of malnutrition amongst children less than 5 years of age in developing countries. However, approximately 27% of children under the age of 5 in these countries are still malnourished. This work focuses on the childhood malnutrition in Gamo Gofa Zone, Ethiopia. This study examined the association between demographic and socioeconomic determinants and the malnutrition problem in children less than 5 years of age using Data obtained from both rural and urban sampled surveys conducted in both Dita and Mirab Abaya Weredas from December 1 to January 5, 2013. The study on the Child undernutrition and underweight prevalence in Gamo Gofa has allowed us to quantify the negative impacts of child undernutrition in both social and economic terms. Today, more than 4 out of every 7 children in Ethiopia are stunted. The result revealed that as many as 75% of all cases of child undernutrition and its related pathologies go untreated. It is also observed that about 35% of the health costs associated with undernutrition occur before the child turns 1 year-old. Generally, The results of the analysis show that place of residence, employment status of mother, employment status of partners, age of the child, educational status of mothers, diarrhea, household economic level and source of drinking water were found to be the most important determinants of health/nutritional status of children. The study revealed that socio-economic, demographic and health and environmental variables have significant effect on the nutritional and health status of children in Ethiopia. The study revealed that children from employed mothers are at a higher risk of health problem and malnutrition.

Keyword: Bayesian Models, Childhood Malnutrition, Gamo Gofa Zone, Ethiopia
Sensitivity analyses for dependent censoring in survival analysis

Alan Kimber, Stefanie Biedermann and Natalie Staplin

There are often reasons to believe that there may be dependence between the time to event and time to censoring in survival data, particularly in a medical context. The motivating example for this study relates to the survival of patients on the waiting list for a liver transplant where, broadly speaking, the most ill patients are prioritised to receive a transplant. Such censoring is known as dependent censoring and is in contrast to the assumption of non-informative censoring that underlies standard survival techniques. Because of identifiability issues, an approach based on sensitivity analysis is used here: a standard survival model that assumes the censoring is non-informative (such as a semi-parametric Cox model or a parametric regression model) is fitted to the survival data as the basis for the analysis and then the effect of small or modest amounts of dependent censoring on the quantity of interest (such as a parameter estimate or a linear predictor) is quantified. In this paper the basic sensitivity analysis methodology is introduced and reviewed, and the approach is illustrated using transplant waiting list data.
Survival Analysis Models for MESS Epilepsy Data

Boryana Cristina Lopez Kolkovska

In the study of medical survival data, such as studies about the recurrence of epileptic seizures in a patient under treatment, the object of interest often lies in the effect that the treatment will have in the seizure recurrence, and the possibility of the patient reaching a state of remission. We propose a Survival Analysis approach to study the risk of seizure recurrence for the Multicentre Study of Early Epilepsy and Single Seizures (MESS). A parametric joint model is proposed which considers that each patient has an underlying unknown seizure rate, subjected to change after taking an antiepileptic drug. A first approach is done by applying a negative binomial mixture model, which assumes that the times to first seizures after randomization to a drug have an exponential distribution, conditional on a frailty term. The model allowed for the possibility that there might be a proportion of the patients that will attain remission without experiencing seizures after the treatment. We perform a residual analysis in order to study the goodness of fit of this model, and compare the fits between the negative binomial mixture model, and the commonly used semiparametric Cox proportional hazards model. We observe the difference in robustness between these two models by taking out of the study a number of outliers and extreme values, and perform a simulated residual study to show how the Cox model performs with data produced from a Poisson process. Further generalizations of the negative binomial model are shown, given that a single seizure is not sufficient to make a diagnosis of epilepsy, a zero-truncated negative binomial model is the natural next step. We finally move on to propose a competing risks model in which the different types of seizures, dropout and death are considered as the competing risks in the population.

Boryana Cristina Lopez Kolkovska
University of Warwick, Coventry, United Kingdom
e-mail: b.c.lopez-kolkovska@warwick.ac.uk
Calculating a population attributable fraction in the case of clustered data

Hendriek Boshuizen and Marit de Lange

The Population Attributable Fraction (PAF) measures the impact of a risk factor on public health. It describes the percentage of disease eliminated when the risk factor would be removed. Traditionally, this measure is calculated at the population level, using survey data on the prevalence of the risk factor, and relative risks from literature. Currently, the PAF is often calculated also directly for a particular epidemiological study. In that case the study delivers both the prevalence of the risk factor as well as the relative risk between risk factor and disease. In this case employing the PAF-formula that is used for population level estimation is circuitous and error-prone. Therefore, a direct approach is preferable. The direct approach uses the statistical model for the disease as a function of the risk factor and other covariates to predict the probability of disease in each individual (and thus also in the entire population) in the case that exposure to the risk factor is eliminated. Together with the observed disease prevalence this allows calculation of the PAF. The direct approach method was applied in a study investigating differences in adverse pregnancy outcome between women residing in Q fever-affected areas and those living elsewhere. For these data the obvious model is a random intercept logistic model. Predicting population level prevalence in such models is more complicated. Using the Best Linear Predictor (BLUP) for each participant yields estimates for individuals that are shrunken to the population mean. Given the nonlinearity of the logistic model this yields biased estimates of the mean population prevalence. Moreover, ML estimates of random effects, as delivered by standard software, are also biased. We compare using BLUPs with proper averaging over the random effects and compare different methods for estimating the random effects (Maximum Likelihood, Restricted Pseudo-Likelihood) using simulation based on our motivating example.
Measurement error models for accelerometer data using MCMC, Integrated Nested Laplace Approximation and Maximum Likelihood Estimation methods

George O. Agogo, Hilko van der Voet, Pieter van’t Veer, Laura Trijsburg, Fred A. van Eeuwijk and Hendriek Boshuizen

In physical activity research, accelerometer is used to measure total energy expenditure (TEE). The accelerometer measurements are, however, subject to measurement error. We quantify measurement error in the accelerometer with correlation coefficient between true and measured TEE (validity coefficient) and an attenuation factor. We use doubly labelled water (DLW) measurements as the gold standard. A DuPLO validation study consisted of 200 individuals (46% male) aged 25-69 years. Among study participants, 30 had two DLW replicate measurements whereas 70 had only one replicate. Similarly, 29 of the participants had two accelerometer replicate measurements whereas 179 had one replicate. We present a bivariate linear mixed measurement error (LMME) model for accelerometer and DLW measurements. The bivariate LMME model contains latent variable for true TEE, bias and random error terms. We estimate the validity coefficient and the attenuation factor using Bayesian Markov Chain Monte Carlo (MCMC), Bayesian Integrated Nested Laplace Approximations (INLA) and frequentist maximum likelihood estimation (MLE) methods. We bootstrap to obtain quantile distribution for the MLE estimate. The posterior mean (95% credible interval) estimate for the validity coefficient with MCMC was 0.770 (0.627; 0.874) and with INLA was 0.794(0.678; 0.882); the point (95% confidence interval) estimate with MLE was 0.810 (0.631; 0.852). Similarly, the posterior mean (95% credible interval) estimate for the attenuation factor with MCMC was 0.786 (0.517; 1.036) and with INLA was 0.813(0.609; 0.992); the point (95% confidence interval) estimate with MLE was 0.818 (0.582; 1.041). The three methods estimate both the validity coefficient and the attenuation factor approximately as 0.8, close to one for error-free measurements. This implies modest attenuation of a regression slope that relates TEE with a given health outcome due to accelerometer measurement error. INLA is computationally efficient but faces difficulty in handling product of two latent Gaussian variables present in the measurement error model.

George Agogo
Wageningen University, Wageningen, Netherlands
e-mail: george.agogo@wur.nl
Dealing with impreciseness in ranked set sampling with fuzzy sets

Bekir Cetintav and Selma Gurler

Ranked set sampling is a useful sampling method which improves the precision of the estimator of the population mean. It uses the ranks of the units instead of the actual measurement whenever the exact measuring is hard and/or expensive. The ranking mechanism is one of the major parts of ranked set sampling procedure. Ranking the units (visually or by a concomitant variable) could not be perfect because there is an ambiguity in discriminating the rank of one unit with another. Clearly, there is an uncertainty for ranking mechanism assigning to ranks and it cause information loss. Fuzzy sets could be a good way to dealing with this uncertainty occurs in the ranking. The most important advantage of using fuzzy sets is to allow the units to belong to different sets with different membership degrees. When we use fuzzy sets, the units in the sets could belong to not only the most possibly rank but also the other possible ranks. In this study, we propose a new sampling procedure, called fuzzy-weighted ranked set sampling (FRSS), with combining the uncertainties coming from the ranking mechanism and observed units by using fuzzy sets concept. We introduce an estimator for the population mean using the fuzzy membership degrees as weights. A simulation study is constructed to evaluate the new method and estimator. Also a biometric real-life data have been used to illustrate the application of the new sampling approach and a cost-benefit analysis has been done based on efficiencies of the methods. The results show that this new method is more efficient than the simple random sampling and ranked set sampling methods for estimation of the population mean and it could be a remarkable alternative for the methods used in biometrics.
Poster lightning presentations P2
Enrichment of GWAS data, combining different sources of genetic data

Erik van Iperen, G.K. Hovingh, F.W. Asselbergs and A.H. Zwinderman

In the past decade many GWAS were performed that discovered many new associations between SNPs and different phenotypes. Imputation methods are widely used in GWAS. They facilitate the association of variants that are not directly genotyped, with phenotypes. Imputation methods can also be used to combine and analyse data genotyped on different genotyping arrays. In this study we investigated if there is a difference in imputation quality between two different approaches of combining GWAS data from different platforms. We will investigate if combining data from different platforms before the actual imputation performs better than combining the data from different platforms after imputation. Methods: Genotype data from the AMC-PAS cohort were used. Samples were genotyped on 3 different platforms. A total of 706 individuals were genotyped on the MetaboChip, 757 individuals on the 50K gene-centric Human CVD BeadChip and 955 individuals on the HumanExome chip. Minimac in combination with MaCH was used for imputation with the 1000genomes reference panel. All imputed markers with an r2 value of <0.3 were excluded in our post-imputation QC. Results: All three datasets were carefully matched on strand, RS-number and genomic coordinates. This resulted 979 unique individuals and a total of unique 258925 markers. A total of 4117036 SNPs were available after combining the three different sets after individually imputation versus a total of 3933494 SNPs after the imputation of the combined set. Our results suggest that the imputation of the different platforms independently performs better than combining the platforms before imputation. We found the opposite result for the analysis where we only had a look at the overlapping individuals on the 3 platforms. Conclusions: Based on the results we concluded that combining the data from three different platforms together after imputation performs better than combining the data of the 3 platforms before imputation.

Erik van Iperen
Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Amsterdam, The Netherlands
e-mail: e.p.vaniperen@amc.uva.nl
Variational Bayesian SEM for undirected Network recovery using external data

Gino Kpogbezan, Mark van de Wiel, Aad van der Vaart, Wessel N. van Wieringen and Gwenael G.R. Leday

Recently we developed a Bayesian structural equation model (SEM) framework with shrinkage priors for undirected network reconstruction. It was shown that Bayesian SEM in combination with variational Bayes is particularly attractive as it performs well, is computationally very fast and a flexible framework. A posteriori variable selection is feasible in our Bayesian SEM and so is the use of shrinkage priors. These shrinkage priors depend on all regression equations allowing borrowing of information across equations and improve inference when the number of features is large. An empirical Bayes procedure is used to estimate our hyperparameters. We also showed in simulations that our approach can outperform popular (sparse) methods. Here, we focus on addressing the problem of incorporating external data and/or prior information into network inference. In many settings information regarding network connectivity is often available. It is then natural to take such information into account during network reconstruction. Based on Bayesian SEM we propose a new model that focuses on the use of external data. It performs better than that of our Bayesian SEM when the external information is relevant, and as good when it is not.
Alternative splice variants in gene expression values in patients with Marfans syndrome

Wouter Ouwerkerk and Aeilko H. Zwinderman

Alternative splicing of messenger RNAs provides cells with the opportunity to create protein isoforms of a multitude of functions from a single gene by excluding or including exons during post-transcriptional processing. Reconstructing the contribution of these splice variants on the total amount of gene expression remains difficult.

Methods We introduced a probabilistic formulation of the alternative splicing reconstruction problem using a finite mixture model, and provide a solution based on the maximum likelihood principle. Our model is based on the assumption that the expected expression level of exons in a particular splice variant is the same for all exons in that variant but allows for measurement error.

In this algorithm the expression in a patient can be written as a weighted sum of the number of splice variant mixture multivariate Gaussian densities. We estimated the model parameter by maximizing the total likelihood using a Nelder and Mead optimization algorithm in R.

To evaluate our algorithm we compared the AIC/BIC values of six models: Established optimal normal mixture modeling method, all exons are equally transcripted, the currently known splice variants, all possible splice variants, the known variants aided with the high prevalent variants of the all possible variants model, and manually selected splice variants.

We applied the models to three genes (SLC2A10, TGFR2 and FBN1), with 25, 29 and 265 possible splice variants, associated with Marfana’s syndrome in gene/exon expression data of 63 patients with Marfans syndrome. The models with the known splice variants aided with the high prevalent splice variants from the all possible splice variants had the best AIC/BIC values for all three genes. In SLC2A10 and FBN1 there was one, in TGFR2 two predominant splice variants. We found four possible new splice variants in three genes associated with Marfans syndrome.

Wouter Ouwerkerk
Academic Medical Center, Amsterdam, Netherlands

e-mail: w.ouwerkerk@amc.uva.nl
Combination approaches improve predictive performance of diagnostic rules for mass-spectrometry proteomic data

Alexia Kakourou, Bart Mertens and Werner Vach

We consider an approach for combining a collection of individual classifiers for the construction of classification rules for mass spectrometry proteomic data, as opposed to calibrating a single predictor only. In a first stage this is achieved by calibrating distinct classifiers separately, each-one using the entire proteomic data set. A double leave-one-out cross-validatory approach is used to estimate the class predicted probabilities on which the combination method will be calibrated. In addition, we propose a new approach to the simulation of proteomic mass spectrometry classification data based on re-use of existing spectrometry data and used this to implement an extensive evaluation of the proposed combination methods. Results from both the data analysis and from the simulation study show that gains in classification performance and predictive accuracy can be achieved with a combination method.
Recent researches show that the sensitivity of cancer cell lines to drugs depends on multiple genomic features of the cell lines. In other words, there is a strong link between mutations in genomic features of the cancer cell lines and the responses to the treatment. The existing methods of statistical modeling of these kind of data are based on simple or multiple regression. In order to improve this analysis, what we consider is fitting a multivariate regression to these data as an alternative method. All these genomic features are considered as covariates and responses of these cancer cells to drugs are the multivariate response variable. As these data are high dimensional, in order to resolve the high dimensionality problem of the data, a new statistic is introduced as a useful tool for screening covariates. The resulted lower dimensional model will only contain the important cancer causing factors.
Sample size determination for classification using metabolomics data


Metabolomics is the simultaneous analysis of hundreds of metabolites present in crude extracts, such as those from plants. This provides detailed insight into metabolite differences and similarities between crude extracts corresponding to different backgrounds. When this background is one of two classes (e.g. control and infected), metabolomics can be used to identify metabolites differentially expressed in these two classes. These metabolites have the potential to function as a biomarker, predicting the state of unknown samples. An important first step for this type of metabolomics studies is the determination of the optimal sample size. If we do not include enough samples of both cases in our study, we may fail to achieve a correct classification and/or have one that could also have been obtained by chance alone. Adding too many experimental units puts unnecessary strains on budgets, etc. In this talk, we demonstrate a simulation approach that estimates optimal sample sizes for the classification of two classes using metabolomics data. Using a pilot data set, a (small) set of most distinguishing metabolites is selected using standard t-tests. For this subset of metabolites, two covariance matrices are estimated, one for each of the two classes. By drawing from a multivariate normal distribution using the two covariance matrices, simulated data sets with varying sample sizes are created for the two classes. These simulated data sets are used in a double cross validated stepwise regression model to predict the class membership. Given the predicted class membership, Hotelingas T2 is calculated. A permutation test is also applied to see if the result of the classification could possibly have been obtained by chance alone. This procedure is repeated many times. For all sample sizes this results in many T2 values (and their significances) and permutation test results. Next, for each sample size, we calculate the fraction significant classification using the T2 tests and the fraction significant permutation tests. The smallest samples sizes where both fractions are at least 90% are considered the optimal sample sizes.

Jos Hageman
Biometris - Applied Statistics, Wageningen University, Wageningen, The Netherlands
e-mail: jos.hageman@wur.nl
The use of whole-genome sequence (WGS) data is expected to enable accurate estimation of effects of actual mutation on a given trait directly, which is highly beneficial compared to current single nucleotide polymorphism (SNP) panels. Despite falling costs of sequencing, sequencing a large number of individuals is still too expensive. A promising approach is to sequence a core set of individuals and use these data to impute missing genotypes for individuals genotyped at lower density. The objectives of the study were 1) to investigate accuracy of imputation from lower density SNP panels to WGS data, and 2) to investigate reliability of genomic prediction based on imputed WGS data. Objective 1 was studied using WGS data of 114 Holstein Friesian bulls. Imputation from a 50k SNP panel and a 777k SNP panel to WGS data was studied. Beagle 3.3.2 software was used for imputation. Objective 2 was studied using 777k SNP panel data and imputed WGS data of 5503 Holstein Friesian bulls. Genomic prediction was performed via a Bayesian stochastic search variable selection method and via a restricted maximum likelihood model that uses a genomic relationship matrix. Accuracy of imputation was generally high (0.77 - 0.83) for imputation from the 777k SNP panel to WGS data, but was low (0.37 - 0.46) for imputation from the 50k SNP panel to WGS data. Sequential imputation from the 50k SNP panel to 777k SNP panel to WGS data substantially improved accuracy of imputation. Minor allele frequency, linkage disequilibrium between a known SNP and imputed SNP, and reference group size affected imputation accuracy. Genomic prediction using the Bayesian method did perform better in all cases. However, no advantage of genomic prediction using WGS data compared to 777k SNP panel data was found. Suggestions to improve accuracy of genomic prediction using WGS data will be discussed.
In Statistical Genetics we are confronted with increasing amounts of genotyping data. To analyze those data, we need appropriate statistical methods as well as ways to visualize them. Here we focus on linkage disequilibrium decay (LD decay). Genetic markers on the same chromosome are not independent. The strength of these correlations decreases with increasing genetic distances between the markers: LD decay. LD decay is typically displayed in a scatterplot of pairwise LD between markers versus marker distance. When the number of markers increases, the interpretation of such scatter plots becomes problematic. With thousands of markers, we get millions of comparisons in terms of correlations. The scatter plot can be improved in several ways: one can fit a (non-) parametric curve to the data cloud to describe the mean relation between LD and marker distance. As an example of the extension of this approach we fit expectile curves. Expectiles give insight into the trend as well as the spread of the data. Computing times are considerably shortened by summarizing the data on a two-dimensional grid in the domain of distance and correlation. For a 100 by 100 grid we get a maximum of $10^4$ pseudo-observations, independent of the number of initial data pairs. A scatter plot smoother also tackles this problem. The resulting information about the mid-points of the bins and the weights can be used for further analysis, e.g. with the expectile or quantile curves. The methods will be illustrated using data from plant genetics.
Ridge estimation of inverse covariance matrices from high-dimensional data

Wessel van Wieringen and Carel Peeters

Molecular biology aims to understand the molecular processes that occur in the cell. That is, which molecules present in the cell interact, and how is this coordinated? For many cellular process, it is unknown which genes play what role. A valuable source of information to uncover gene-gene interactions are (onco)genomics studies. Such studies comprise samples from n individuals with, e.g., cancer of the same tissue. Each sample is interrogated molecularly and the expression levels of many (p) genes are measured simultaneously. From these high-dimensional omics data the gene-gene interaction network may be unravelled when the presence (absence) of a gene-gene interaction is operationalized as a conditional (in)dependency between the corresponding gene pair. Then, under the assumption of multivariate normality, the gene-gene interactions correspond to zero’s in the precision matrix (which are proportional to the partial correlations). When dealing with high-dimensional data, the sample covariance matrix is singular and the sample precision matrix is not defined. But even if $p > n$ and $p$ approaches $n$, the sample precision matrix yields inflated partial correlations. Both situations require a form of regularization to obtain a well-behaved estimate of the precision matrix, and consequently of the gene-gene interaction network. We study ridge estimation of the precision matrix in the high-dimensional setting. We first review two archetypal ridge estimators and note that their penalties do not coincide with common quadratic ridge penalties. Subsequently, starting from a proper $\ell_2$-penalty, analytic expressions are derived for two alternative ridge estimators of the precision matrix. The alternative estimators are compared to the archetypes with regard to eigenvalue shrinkage and risk. The alternatives are also compared to the graphical lasso within the context of graphical modeling. The comparisons may give reason to prefer the proposed alternative estimators.

Wessel van Wieringen
VU University Medical Center, Amsterdam, Netherlands
e-mail: w.vanwieringen@vumc.nl
Satellite meeting
Inferring dynamic genetic networks
We propose a chain graphical model for high dimensional time course data with the aim to reconstruct the structure and dynamics of processes on networks. The proposed model is parameterized by a precision and autoregressive coefficient matrices to explore patterns of contemporaneous and dynamic interactions. We use a penalized likelihood inference that efficiently combines the idea of Gaussian graphical models and Bayesian dynamic networks using a smoothly clipped absolute deviation and lasso penalties. We obtain sparse estimates that account for highly structured systems. Moreover, we propose a dynamic chain graphical model that borrows strength across time. The method is illustrated on simulated data and time course microarray gene expressions data from Arabidopsis Thaliana and mammary gland development.
Sparse vector autoregressive models with an application in marketing

Ines Wilms

When there are few data relative to the model complexity, standard estimation techniques lack power to identify existing relationships and make accurate forecasts. We introduce a sparse estimation technique for the typically heavy parameterized vector autoregressive market response model. The method is used to identify demand effects in a large network of product categories. Using this sparse estimation procedure, we obtain a parsimonious product category network without prior constraints on the structure.

____________________________
Ines Wilms
Katholieke Universiteit Leuven, Leuven, Belgium
e-mail: ines.wilms@kuleuven.be
A focused selection of graphical models in fMRI connectivity studies

Gerda Claeskens

We develop a focused information criterion for graphical models to determine brain connectivity. The proposed method can be tailored to specific research questions. It selects a graph with a small estimated mean squared error for a user-specified focus, which allows to construct individual-tailored graphs. The graphs may be of small size, or may contain more nodes than the number of samples in which case penalized estimation methods are used. The graphical models may include autoregressive times series components, they can relate graphs from different subjects, or pool data via random effects. Examples illustrate the method.
Controversy in mechanistic modelling of biopathways with Gaussian processes

Dirk Husmeier

Many processes in biology can be described by dynamical systems models based on ordinary differential equations (ODEs). Examples range from simple models of autocatalysis in chemical kinetics or activation/deactivation dynamics of spiking neurons to increasingly complex mathematical descriptions of biopathways that aim to predict the time-varying concentrations of different molecular species, like mRNAs and proteins, inside the living cell. However, parameter inference and structure learning in mechanistic models based on non-affine differential equations is computationally onerous due to the need for numerical integration of the ODEs. For that reason, various faster alternatives based on gradient matching have been proposed in the recent literature. A particularly promising approach is based on nonparametric Bayesian modelling with Gaussian processes, which exploits the fact that a Gaussian process is closed under differentiation. Two alternative paradigms have been proposed. The first paradigm is based on a product of experts approach and a marginalization over the derivatives of the state variables. The second paradigm is based on a probabilistic generative model and a marginalization over the state variables. The claim has been made that this leads to better inference results. In my talk, I will offer a new interpretation of the second paradigm, which highlights the underlying assumptions, approximations and limitations more clearly. I will present theoretical and empirical studies to assess the viability of both paradigms, compare their effectiveness with the alternative framework of emulation, and discuss the prospects for structure learning in mechanistic models of biopathways.

Dirk Husmeier
University of Glasgow, Glasgow, UK
e-mail: Dirk.Husmeier@glasgow.ac.uk
Realistic network reconstruction methods for the Timing Metabolism (TiMet) research project

Marco Gregorzyk

We extend a recently proposed gradient-matching method for inferring interactions in complex systems described by differential equations in various respects: improved gradient inference, evaluation of the influence of the prior on kinetic parameters, comparative evaluation of two model selection paradigms: marginal likelihood versus DIC (divergence information criterion), comparative evaluation of different numerical procedures for computing the marginal likelihood, extension of the methodology from protein phosphorylation to transcriptional regulation, and a critical model evaluation based on a realistic simulation of the underlying molecular processes with Markov jump processes.

Marco Gregorzyk
University of Groningen, Groningen, the Netherlands
e-mail: m.a.grzegorczyk@rug.nl
Graphical models for causal inference from multivariate time series

Michael Eichler

In time series analysis, inference about cause-effect relationships among multiple time series is commonly based on the concept of Granger causality, which exploits temporal structure to achieve causal ordering of dependent variables. One major and well known problem in the application of Granger causality for the identification of causal relationships is the possible presence of latent variables that affect the measured components and thus lead to so-called spurious causalities. We present a new graphical approach for describing and analysing Granger-causal relationships in multivariate time series that are possibly affected by latent variables. It is based on mixed graphs in which directed edges represent direct influences among the variables while dashed edges—directed or undirected—indicate associations that are induced by latent variables. We show how such representations can be used for inductive causal learning from time series and discuss the underlying assumptions and their implications for causal learning. Finally we will discuss tetrad constraints in the time series context and how the can be exploited for causal inference.

Michael Eichler
Maastricht University, Maastricht, the Netherlands
e-mail: m.eichler@maastrichtuniversity.nl
List of Presenters

· Fentaw Abegaz: Satellite
· Magne Aldrin: Contributed C10
· George Agogo: Poster P1
· Eleni-Rosalina Andrinopoulou: Contributed C8
· Tilahun Ferede Asena: Poster P1
· Julie Aubert: Contributed C13
· Leyla Azarang: Contributed C9
· Elleny Balder: Poster P1
· Theodor Adrian Balan: Contributed C9
· Johannes Berkhof: Poster P1
· Rianne van Binsbergen: Poster P2
· Paul Blanche: Invited 1
· Hendriek Boshuizen: Poster P1
· Martin Boer: Contributed C7
· Said el Bouhaddani: Contributed C1
· Cajo ter Braak: Contributed C5
· Susan Bryan: Contributed C12
· Bekir Cetintav: Poster P1
· Gerda Cleaskens: Satellite
· Diana Cole: Contributed C10
· Daniel Commenges: Contributed C14
· Fred van Eeuwijk: Contributed C13
· Michael Eichler: Satellite
· Paul Eilers: Contributed C13
· Nicole Erler: Contributed C12
· Marie-Pierre Etienne: Invited 2
· Pablo Ferrer: Contributed C8
· Koen Fussenich: Contributed C3
· Paul Garthwaite: Contributed C3
· Nan van Geloven: Contributed C11
· Maria Gheorghe: Contributed C5
· Robert Goudie: Contributed C5
· Jos Hageman: Poster P2
· Jesse Hemerik: Contributed C15
· Tom Heskes: Invited 3
· John Hinde: Contributed C4
· Michel Hof: Contributed C4
· Dirk Husmeier: Satellite
· Erik van Iperen: Poster P2
· Anna Ivanova: Contributed C12
· Rianne Jacobs: Contributed C10
· Monika Jelizarow: Contributed C6
· Natoya Jourdain: Contributed C10
· Alexia Kakourou: Poster P2
· Alan Kimber: Poster P1
· Mia Klinten Grand: Contributed C11
· Marco Gregorzyk: Satellite
· Philippe Lambert: Contributed C5
· Boryana Cristina Lopez Kolkovska: Poster P1
· Gino Kpogbezan: Poster P2
· Andy Lynch: Contributed C1
· Ivonne Martin: Contributed C2
· Chamberlain Mbah: Contributed C7
· Renee de Menezes: Contributed C1
· Viktoriaan Miok: Contributed C2
· Jammbe Musoro: Contributed C11
· Jan De Neve: Contributed C9
· Elaheh Oftadeh: Poster P2
· Wouter Ouwerkerk: Poster P2
· Carel Peeters: Contributed C2
· Hein Putter: Contributed C9 & Invited 1
· Sylvia Richardson: Keynote
· Dimitris Rizopoulos: Invited 1
· Mar Rodriguez: Contributed C6
· Anais Rouanet: Contributed C8
· Michael Schaub: Invited 2
· Sabine Schnabel: Poster P2
· Aldo Solari: Contributed C15
· Simon Thompson: Contributed C14
· Sophie Vanbelle: Contributed C3
· Sjouke Vandenberghe: Contributed C7
· Nadia Vendrig: Contributed C13
· Roel Verbelen: Contributed C9
· Pablo Verde: Contributed C5
· Karel Vermeulen: Invited 3
· Matthijs Vynck: Contributed C1
· Ingeborg Waernbaum: Invited 3
· Junfeng Wang: Poster P1
· Mathilde Wanneveich: Contributed C4
· Ron Wehrens: Contributed C6
· Alan Welsh: Contributed C15
· Wessel van Wieringen: Poster P2
· Ines Wilms: Satellite
· Ernst Wit: Contributed C14
· Birgit Witte: Contributed C9
· Heather Whitaker: Poster P1
· Yalcin Yavuz: Contributed C15
· Chaozhi Zheng: Contributed C1