# 21st Young Statisticians Meeting
# YSM 21

PROGRAMME — ABSTRACTS — PARTICIPANTS

NOVEMBER 4-6, 2016
PIRAN, SLOVENIA


SUPPORTED BY

CENTER FOR METHODOLOGY AND INFORMATICS, FDV, UNIVERSITY OF LJUBLJANA
NIB – NATIONAL INSTITUTE OF BIOLOGY
STATISTIČNO DRUŠTVO SLOVENIJE

**Local Organizer**

Anuška Ferligoj
University of Ljubljana, Faculty of Social Sciences
Kardeljeva pl. 5, 1000 Ljubljana, Slovenia
Anuska.Ferligoj@fdv.uni-lj.si

**International Program Committee**

Austria: Herwig Friedl, Graz
Croatia: Ksenija Dumičić, Zagreb
Hungary: Tamás Rudas and Renata Nemeth, Budapest
Italy: Dario Gregori and Paola Berchialla, Padova
Slovenia: Anuška Ferligoj, Ljubljana

Published by: CMI, FDV, University of Ljubljana

Edited by: Vladimir Batagelj and Anuška Ferligoj

Logo Design by: Andy Posner

# Programme

# Programme of the 21st Young Statisticians Meeting

## Friday, November 4, 2016

17:30 – 19:00 Registration at Marine Biology Station

## Saturday, November 5, 2016

8:00 – 9:00 Registration at Marine Biology Station

9:15 – 9:30 Welcome address

9:30 – 11:00 **Session 1** Chair: T. Rudas
***Sandor Pecsora*** (Hungary): A general approach to probabilistic inequalities
***Rok Okorn*** (Slovenia): Lévy modeled GMWB: Pricing with wavelets
***Beáta Bolyog*** (Hungary): Statistical inference for two factor affine diffusions

11:00 – 11:30 **COFFEE BREAK**

11:30 – 13:00 **Session 2** Chair: J. Stare
***Črt Ahlin*** (Slovenia): Restricted cubic splines for periodic data
***Corrado Lanera*** (Italy): Bayesian Machine Learning Techniques for revealing complex interactions among genetic and clinical factors in association with extra-intestinal manifestations in IBD patients
***Christine Wallisch*** (Austria): External validation of cardiovascular risk prediction models in the Austrian health screening population

13:00 – 14:00 **LUNCH**

14:00 – 15:30 **Session 3** Chair: P. Berchialla
***Sercan Gür*** (Austria): Pricing Parisian option with adaptive Monte Carlo Method
***Maja Zagorščak*** (Croatia): Iterative motif scanning in HMM framework
***Jelena Kovačić*** (Croatia): Developing prediction models using a small number of datasets with overlapping variables

15:30 – 16:00 **COFFEE BREAK**

16:00 – 17:30 **Session 4** Chair: K. Dumičić

***Kitti Balogh*** (Hungary): Unveiling latent topic structure in anti-Roma discourse using Latent Dirichlet Allocation

***Rita Cimmino*** (Italy): Scientific collaboration network in Italian community: A mixed method approach

***Jasmina Pivar*** (Croatia): SOM Ward clustering approach for client segmentation in leasing industry

19:00 **DINNER**

## Sunday, November 6, 2016

9:30 – 11:00 **Session 5** Chair: H. Friedl

***Philipp Hermann*** (Austria): Estimating variable recombination rates

***Danila Azzolina*** (Italy): Handling missing or incomplete data in a Bayesian network meta-analysis framework

***Klemen Pavlič*** (Slovenia): Using pseudo observations for estimation of net survival

11:00 **CLOSING**

# Abstracts

# Restricted cubic splines for periodic data

**Črt Ahlin** (Slovenia)
e-mail: crt.ahlin@gmail.com

Many biomedical outcome variables display seasonal or periodic variations. For example, one might be interested in modelling the probability of isolating disease-causing bacteria from samples obtained from patients diagnosed over several years. A non-monotonic association between the day of the year when the patients were diagnosed and the probability of obtaining a positive result can be expected for the diseases where the bacteria are transmitted from seasonally active vectors. Similarly, some viruses are detected with known patterns of seasonality.

Restricted cubic splines (RCS) are commonly used in regression to model nonlinear associations between explanatory variables and outcomes, without requiring the specification of the exact form of nonlinearity. In practice, the nonlinearity of the association is tested using standard hypothesis testing methods and the estimated shape of the association is displayed graphically. For periodic variables the use of RCS is not an optimal choice, as they do not take the periodic nature of the variables explicitly into account. For example, the naive use of RCS does not guarantee the equality of the estimated outcomes at the beginning and at the end period.

In this talk we derive the basis functions of periodic RCS (per-RCS); the models with per-RCS are extremely parsimonious, as they require the estimation of only k-3 parameters for splines with k knots. We compare per-RCS to RCS and to periodic cubic splines (per-CS), which were previously proposed for modelling periodic data. Using real and simulated data we show that per-RCS outperform per-CS, obtaining more precise and generalizable estimates and better calibrated models. The association between the covariate and the outcome is detected with higher statistical power, while maintaining the correct size of the statistical tests and coverage of the point-wise confidence intervals.

# Handling missing or incomplete data in a Bayesian network meta-analysis framework

**Danila Azzolina** (Italy), Ileana Baldi, Clara Minto, Dario Gregori
e-mail: danila.83@live.com

A Bayesian NMA model is often used to estimate the effect of each intervention compared to others in the network and the results may be synthesized in terms of rank probabilities of considered treatments. In several cases, a NMA is associated to a loss of information due to incomplete data of studies retrieved through a systematic review, which are therefore excluded from the analysis. Several methods are provided in literature to handle missing or incomplete data in a NMA framework.

It is often the case that only baseline and follow-up measurement are available; to obtain data about mean change it is necessary to consider pre-post study correlation. In a Bayesian framework, some authors (Abrams, 2005), suggest imputation strategies of pre-post correlation by generating a posterior distribution of Fisher's transformation to obtain a posterior estimate of correlation.

In other cases, a variability measure associated to mean change score might be unavailable. Different imputation methods are suggested in literature, as those based on maximum standard deviation imputation. The purpose of this study is to verify the robustness of Bayesian NMA models with respect to different imputation strategies through simulations.

Fifty trials are simulated in full databases by including baseline, follow-up and Delta variation information. Baseline data are obtained by sampling from bounded 0-100 normal distributions (X N(41.8,21.5)) (Cannon, 2000), to mimick the support of WOMAC score. Delta variation data are simulated from normal distributions with parameters provided by a literature review about 6 FANS treatments against symptoms of knee osteoarthritis as shown above.

| | Mean delta | SD Delta | Author-Year |
|---|---|---|---|
| Rofecoxib | -26.70 | 22.99 | Cannon, 2000 |
| Diclofenac | -29.60 | 23.39 | Cannon, 2000 |
| Etoricoxib | -31.03 | 20.13 | Reginster, 2006 |
| Licofelone | -18.18 | 25.88 | Raynauld, 2008 |
| Naproxen | -19.31 | 25.68 | Raynauld, 2008 |
| Placebo | -9.94 | 21.29 | Kahan, 2009 |

Follow up variability data are provided from generated Delta and baseline variability measures, using inverse formula, setting hypotheses on pre-post correlation and considering, in each scenario, a sequence from 0.3 to 0.95 by 0.05.

Sample size are obtained by sampling from an uniform distribution bounded 50-100. Between trial heterogeneity has been included as a variability measure by following, for each simulation setting, a sequence from 0.1 to 5 by 0.1. Each simulation scenario provides different combination of heterogeneity between trials and pre-post correlation creating 700 scenarios.

For each scenario 2 imputed databases, useful to perform a Bayesian NMA, are generated. In the first case, information about Delta variation are randomly removed, from full database, leaving only baseline e follow up data, then variability of mean change has been imputed using correlation method. In another case, also information at baseline and follow up are removed, then Delta variability has been imputed with maximum standard deviation method. Considering every simulated dataset, a multiple treatment meta-analysis, with random effect and Uniform prior on heterogeneity parameter, has been performed using an arm based approach. The Bayesian estimate is based on a MCMC method (200000 iteration, 4 chains) and convergence has been evaluated by Gelman and Rubin diagnostic as indicated in literature.

To investigate robustness of conventional consistency NMA, under several heterogeneity-correlation scenarios and different imputation methods, the bias of rank probabilities estimates has been computed in order to check models performance in ranking treatments accurately. For each scenario, the mean, bias and the standard deviation of the first rank probability estimates, for full and imputed databases, have been computed.

The results show that the bias is very small for every scenario, confirming that ranking provided by models is robust respect to different imputation methods in several heterogeneity-correlation settings. When considering standard deviation, the values are less than 0.0072, but are smallest for models

estimated in a less than one heterogeneity scenarios.

Bayesian NMA seems to be a robust method to rank treatments when incomplete data are imputed. This method is more robust to missing data imputation in a low heterogeneity framework, especially if trials considered in a NMA are conducted on populations sharing similar characteristics.

## References

1. Ades AE, Lu G, Higgins JPT. The interpretation of random-effects meta-analysis in decision models. Medical Decision Making 2005; 25:646–654.
2. Abrams, K. R., Gillies, C. L., & Lambert, P. C. (2005). Meta-analysis of heterogeneously reported trials assessing change from baseline. Statistics in medicine, 24(24), 3823-3844.
3. Cannon, G.W., et al., Rofecoxib, a specific inhibitor of cyclooxygenase 2, with clinical efficacy comparable with that of diclofenac sodium: results of a one-year, randomized, clinical trial in patients with osteoarthritis of the knee and hip. Rofecoxib Phase III Protocol 035 Study Group. Arthritis Rheum, 2000. 43(5): p. 978-87.

# Unveiling latent topic structure in anti-Roma discourse using Latent Dirichlet Allocation

**Kitti Balogh** (Hungary)
e-mail: kttblgh@gmail.com

From the mid 2000's the number of anti-Roma and racist utterances have been increasing in Hungary and this manner of speech has also become accepted in common discourse. The research focused on extracting anti-Roma topics over this period using a hierarchical Bayesian model called Latent Dirichlet Allocation (LDA). The source of the analysis was collected from kuruc.info online newsportal which is the flagship of the far-right media in Hungary. The corpus consists of more than 10.000 anti-Roma news from 2006 until 2015. 27 anti-Roma topics were extracted by using LDA which opens the possibility to analyze the distribution of various topics over time and see how they are connected to the most influential events during the period of investigation. The identified topics correspond to categories identified by qualitative studies on Roma media representation in Hungary. Our research suggests that topic modeling could be a useful supplementary tool to the toolbox of traditional qualitative discourse analysis researchers. Our research project culminated into an interactive data visualization and a data visualization dashboard which can be accessed on following links:

http://labs.precognox.com/kuruc-info-visualization/index_en.htm

http://labs.precognox.com/kuruc-info-dashboard/index.html

# Statistical inference for two factor affine diffusions

**Beáta Bolyog** (Hungary)
e-mail: bbeata@math.u-szeged.hu

We consider general two factor affine diffusions, for $t \in [0, \infty)$

$$\begin{cases} \mathrm{d}Y_t = (a - bY_t)\,\mathrm{d}t + \sigma_1\sqrt{Y_t}\,\mathrm{d}W_t, \\ \mathrm{d}X_t = (\alpha - \beta Y_t - \gamma X_t)\,\mathrm{d}t + \sigma_2\sqrt{Y_t}\,(\varrho\,\mathrm{d}W_t + \sqrt{1-\varrho^2}\,\mathrm{d}B_t) + \sigma_3\,\mathrm{d}L_t \end{cases} \tag{1}$$

where $a \in [0, \infty)$, $b, \alpha, \beta, \gamma \in \mathbb{R}$, $\sigma_1, \sigma_2, \sigma_3 \in [0, \infty)$, $\varrho \in [-1, 1]$ and $(W_t, B_t, L_t)_{t\in[0,\infty)}$ is a 3-dimensional standard Wiener process. In this paper we study asymptotic properties of least squares estimators (LSE) of the drift parameters $(a, b, \alpha, \beta, \gamma)$ based on continuous time observations $(Y_t, X_t)_{t\in[0,T]}$ with $T > 0$, starting the process $(Y, X)$ from some known non-random initial value $(y_0, x_0) \in [0, \infty) \times \mathbb{R}$. It will turn out that for the calculation of the LSE of $(a, b, \alpha, \beta, \gamma)$, one does not need to know the values of the diffusion coefficients $\sigma_1, \sigma_2, \sigma_3$ and $\varrho$. The aim of the present paper is to achieve analogous results to those of Barczy and Pap [2] for the processes given in equations (1), where the maximum likelihood estimator (MLE) is studied for the case of $a, \sigma_1, \sigma_2 \in (0, \infty)$, $\gamma = 0$, $\varrho \in (-1, 1)$ and $\sigma_3 = 0$. We distinguish three cases: subcritical (also called ergodic), critical and supercritical. In the subcritical case, asymptotic normality is proved for all the parameters, while in the critical and supercritical cases, non-standard asymptotic behavior is described. We also compare our results with those of Barczy at al. [1], where the LSE is studied for the case of $a, b, \sigma_1, \sigma_2 \in (0, \infty)$, $\gamma = 0$ and $\sigma_3 = 0$.

**References:**

1 Barczy, M., Nyul, B. and Pap, G. (2016). Least squares estimation for the subcritical Heston model based on continuous time observations. Available at arXiv http://arxiv.org/abs/1511.05948

2 Barczy, M. and Pap, G. (2016). Asymptotic properties of maximum-likelihood estimators for Heston models based on continuous time observations. Statistics 50(2) 389-417.

3 Bolyog, B. and Pap, G. (2016). Conditions for stationarity and ergodicity of two-factor affine diffusions. Submitted.

# Scientific Collaboration Network in Italian Community: A Mixed Method Approach

**Rita Cimmino** (Italy), Anja Žnidaršič, Giancarlo Ragozini, Anuška Ferligoj
e-mail: ritacimmino@gmail.com

In the social sciences, the methodological debates about the opportunity for mixing methods in Social Network Analysis (SNA) are seen in the last years. Despite huge developments in quantitative approaches in SNA more and more researchers have studied social networks by qualitative methods. In this prospective, we consider network as a "social world" of shared meanings, feelings, conventions, norms, and identities. The qualitative methods allow understanding the process of creating of social relations. For this reason, networks are both structure and process at the same time and mixed methods can be very useful. This paper presents the results of a research in an Italian scientific community, adopting a mixed approach for network data collection. The focus is on how mixed methods aid to understanding the scientific collaboration. On one hand, we analyse the co-authorship network to detect the formal collaboration among the authors and, on another hand, we used in-depth interviews to understand informal collaboration and cover other formal types of collaboration (e.g,, co-membership in editorial board, mentorship). The goal is to show how a mixed approach can detect in depth the dynamics related to the scientific collaboration network. Mixed approach enables to map and measure but also to explore issues related to construction, reproduction, variability and dynamism of ties.

# Pricing Parisian Option with Adaptive Monte Carlo Method

**Sercan Gür** (Austria)
e-mail: sguer@wu.ac.at

A Parisian option is a type of barrier option, which can only be exercised if the underlying value process not only reaches a barrier level but remains a certain prescribed time (so-called *window period*) below (or above) this level. Closed form solutions for the value of these contracts do not exist. In order to price Parisian options, we use Monte Carlo simulation instead of partial differential equations, inverse Laplace transform or lattices. We propose a new Monte Carlo method which can be used to price Parisian options not only with constant boundary but with more general boundary. The advantage of this approach is that it can easily be adapted to compute the price of an option with more complicated path-dependent payoff. We use adaptive control variable to improve the efficiency of the Monte Carlo estimator. At last, we provide a numerical example to illustrate our method and a comparison of previous Monte Carlo methods with our technique.

# Estimating Variable Recombination Rates

**Philipp Hermann** (Austria), Andreas Futschik
e-mail: philipp.hermann@jku.at,
e-mail: andreas.futschik@jku.at

Recombination is a natural process in meiosis which increases genetic variation by producing new haplotypes. Populations with higher recombination rates are seen to be more flexible to adapt to new environments. Estimating (population) recombination rates is important in order to understand the process of recombination itself and to localize signals of selection. Recombination rates are heterogeneous between species and also across the DNA sequence of a population.

Common methods for estimating the variable recombination rate as a function of the DNA position use Bayesian approaches. More specifically, a composite likelihood is used within a reversible jump MCMC framework. Two software packages, `LDHat` (Auton 2007) and `LDHelmet` (Chan 2012) are available for this purpose. An improvement of local recombination rate estimators can be obtained via optimizing the trade-off between bias and variance (Gärtner 2016).

Since genome-wide estimations of recombination rates with above mentioned methods are very time consuming, we consider a fast estimation method in our current work. We estimate local recombination rates with relevant summary statistics as explanatory variables in a regression model and yield a much faster estimate than the composite likelihood approach. To estimate locally varying recombination rates we apply a frequentist segmentation algorithm with type I error control (Futschik, 2014).

**References:**

1. Auton, A. and McVean, G. (2007). Recombination rate estimation in the presence of hotspots. Genome Research, 17(8):1219–1227.
2. Chan, A.H., Jenkins, P.A., and Song, Y.S. (2012). Genome-Wide Fine-Scale Recombination Rate Variation in Drosophila melanogaster. PLoS Genetics, 8(12):e1003090.
3. Futschik, A., Hotz, T., Munk, A., and Sieling, H. (2014). Multiscale DNA partitioning: Statistical evidence for segments. Bioinformatics, 30(16):2255–2262.
4. Gärtner, K. and Futschik, A. (2016). Improved Versions of Common Estimators of the Recombination Rate. Journal of Computational Biology.

# Developing prediction models using a small number of datasets with overlapping variables

**Jelena Kovačić** (Croatia)
e-mail: jkovacic@imi.hr

Using multiple data sources to develop clinical prediction models increases sample size and precision. However, when some datasets include only a part of the relevant predictors, a common regression analysis cannot be applied unless one part of the data is discarded. To overcome this issue, a recent study proposed to estimate a regression coefficient from a model with all relevant predictors (fully adjusted estimate, available from at least one dataset) using the correlations and conditional independencies between fully and partially adjusted estimates. To validate the proposed method for the prediction of risk of allergic diseases in Croatian population using 4 datasets, we consider the problem of developing a prediction model when the number of datasets is too small to estimate these correlations reliably. The proposed method, modified to include plausible correlation values in advance, was compared to the complete-case estimator in a simulation study. Although both approaches showed similarly low bias, the mean squared error of the complete-case estimator was larger. These results suggest that the proposed method may be better suited for population prediction models even when the number of datasets is small.

# Bayesian Machine Learning Techniques for revealing complex interactions among genetic and clinical factors in association with extra-intestinal Manifestations in IBD patients

**Corrado Lanera** (Italy), Elisa Menti, Corrado Lanera, Giulia Lorenzoni, Daniela Giachino, Mario de Marchi, Dario Gregori, Paola Berchialla and Piedmont Study Group on the Genetics of IBD
e-mail: corrado.lanera@unipd.it

The objective of the study is to assess the predictive performance of three different techniques as classifiers for extra-intestinal manifestations in 152 patients with Crohn's disease. Naïve Bayes, Bayesian Additive Regression Trees and Bayesian Networks implemented using a Greedy Thick Thinning algorithm for learning dependencies among variables and EM algorithm for learning conditional probabilities associated to each variable are taken into account. Three sets of variables were considered: (i) disease characteristics: presentation, behavior and location (ii) risk factors: age, gender, smoke and familiarity and (iii) genetic polymorphisms of the NOD2, CD14, TNFA, IL12B, and IL1RN genes, whose involvement in Crohn's disease is known or suspected. Extra-intestinal manifestations occurred in 75 patients. Bayesian Networks achieved accuracy of 82% when considering only clinical factors and 89% when considering also genetic information, outperforming the other techniques. CD14 has a small predicting capability. Adding TNFA, IL12B to the 3020insC NOD2 variant improved the accuracy.

# Lévy Modeled GMWB: Pricing with Wavelets

**Rok Okorn** (Slovenia)
e-mail: okorn.rok@gmail.com

Lévy models are increasingly popular in the study of financial and actuarial products. The development of new numerical approaches and the growth of computer power enable practitioners to enjoy better accuracy of the so obtained results. In this presentation we show the study of the problem of a GMWB rider pricing by the means of American option pricing in the Lévy setting. Using wavelet discretization we reduce the problem to a matrix linear complementarity problem giving the value function in the log price domain. In particular, we give the numerical solution of the insurance fee for the rider driven by geometric Brownian motion, Merton jump diffusion and variance gamma process. Our approach with wavelets seems to be helpful in allowing practitioners to compute their fees for this rider as accurately as they wish relatively fast.

# Using pseudo observations for estimation of net survival

**Klemen Pavlič** (Slovenia)
e-mail: klemen.pavlic@mf.uni-lj.si

The field of relative survival analysis has emerged from the need for estimation of burden of a chronic disease in the presence of deaths from other causes. Net survival is the most commonly used measure for comparison of burden of cancer between different populations. It is defined as the average of ratios of individual overall survival and individual population survival. When the data are complete, the individual overall survival can be estimated using indicator function and when the data are right censored, we can use pseudo observations. They enable us to estimate individual overall survival regardless of the censoring and to estimate net survival by directly following its definition. We show that the estimator constructed directly from the definition of net survival has similar properties as the Pohar Perme estimator (an existing estimator with desirable theoretical properties). This also enable us to distinguish between the properties of the Pohar Perme estimator (that has a more complex form) and the properties of the measure. We show how the variance of the new estimator can be expressed and estimated. We derive theoretical formulae and attempt to avoid the overly intensive calculations by using approximations. We compare the precise and approximate approach via simulations. To conclude we will look at the possible extensions of the new estimator to discretely measured data.

# A general approach to probabilistic inequalities

**Sandor Pecsora** (Hungary), Istvan Fazekas, Faculty of Informatics, University of Debrecen
e-mail: pecsora.sandor@inst.unideb.hu

General theorems are obtained on exponential and Rosenthal's inequalities and on complete convergence. The random variables (r.v.'s) $X_1$, $X_2$, ...,$X_n$ are said to be acceptable if

$$\mathbb{E}e^{\sum_{i=1}^{n}\lambda X_i} \leq \prod_{i=1}^{n}\mathbb{E}e^{\lambda X_i} \tag{1}$$

for any real number $\lambda$, see [1]. In this paper we shall show that an appropriate version of inequality (1) implies an exponential inequality. Then the exponential inequality implies a Rosenthal's inequality. Moreover, the exponential inequality implies immediately complete convergence. We emphasize that to obtain the above results no additional dependence conditions are needed. Then our general theorems will be applied to weakly orthant dependent [4] sequences. We mention that the methods of the proofs are closely related to the ones used for independent r.v.'s [2], [3].

**References:**

1 ANTONINI, R.G.; KOZACHENKO, Y.; VOLODIN, A.: Convergence of series of dependent '-sub-Gaussian random variables. J. Math. Anal. Appl. 338 2008, No. 2, 1188-1203.

2 FUK, D. H.; NAGAEV, S. V.: Probabilistic inequalities for sums of independent random variables. Teor. Verojatnost. i Primenen. 16 1971, 660-675.

3 PETROV, V. V.: Limit theorems of probability theory. Sequences of independent random variables. The Clarendon Press, Oxford University Press, New York, 1995.

4 WANG, KAIYONG; WANG, YUEBAO; GAO, QINGWU.: Uniform asymptotics for the
nite-time ruin probability of a dependent risk model with a constant interest rate. Methodol. Comput. Appl. Probab. 15 2013, no. 1, 109-124.

# SOM Ward clustering approach for client segmentation in leasing industry

**Jasmina Pivar** (Croatia)
e-mail: jpivar@efzg.hr

Leasing companies seek for better ways of reaching the clients and improving the effectiveness of their campaigns. One way to do this is to target potential clients with the particular attributes. Market segmentation can be done by using different approaches based on defined several criteria. For example, leasing company may divide the larger market into segments by using demographic and operational criteria. Established criteria may include client's size and address, leasing object, the amount of rent, reference interest rate and so on. Continuously increasing amounts of data in databases are providing companies with the opportunity to gain insight into customer behaviour and predict future agreement status, for example, whether the agreement will be expiry regularly, or there is a possibility for a fraud. The purpose of this study is to determine clients segments, common client profiles and identify segments in which frauds are most possible. SOM-Ward clustering done by using Viscovery SOMine is a useful tool for cluster analysis. We are going to adopt SOM-Ward clustering method to segment the customer base of a Croatian leasing company. The used database contains data on all leasing agreements that were active or completed at the time of running the report, including more than 40 thousands records of raw data. This study proposes to segment the client base according to the demographic and behavioural characteristics of the clients, and operational characteristics related to the leasing agreement.

# External Validation of Cardiovascular Risk Prediction Models in the Austrian Health Screening Population

**Christine Wallisch** (Austria), Georg Heinze, Gerald Mundigler, Wolfgang C. Winkelmayer, Daniela Dunkler
e-mail: christine.wallisch@meduniwien.ac.at

**Introduction**. Austria's health screening program was designed to promote prevention of various diseases including cardiovascular disease (CVD), which is still the single leading cause of death and disability. We assessed if four widely applied and recommended prognostic prediction models with different definitions of CVD - Framingham 1991 and 2008 general CVD models, ACC/AHA 2013 atherosclerotic CVD model, and SCORE 2003 CVD death model - are transportable to the Austrian screening population.

**Methods**. The validation cohort comprised 1.5M individuals participating in the screening program from 2009-2014, aged 30-79 years and without documented CVD history. CVD events were defined by a cardiovascular cause of hospitalization or death. All models were evaluated for a prediction horizon of 5 years by assessing discrimination using Uno's c-index, calibration and clinical utility using decision curves.

**Results**. C-indices ranged from 0.70 to 0.78 and were slightly lower than c-indices obtained in the derivation cohorts. In general, the SCORE model achieved the lowest discrimination. In accordance with other validation studies, c-indices in women were always higher than in men. Calibration curves showed that the Framingham 2008 and the SCORE model overestimated CVD risk in the Austrian screening population, whereas the ACC/AHA model underestimated the observed risk. The Framingham 1991 model obtained good calibration, especially for individuals up to 60 years. Decision curves for the general CVD models and for the atherosclerotic CVD model indicate a net benefit of applying the models when setting the threshold probability at 0.1 or lower.

**Discussion**. C-indices were rather satisfying, except for the SCORE model. Risk stratification for CVD can be accomplished using prediction models in the Austrian screening cohort. However, one might consider recalibration of the prediction models. The Framingham 1991 model seemed to be the most valid, whereas the SCORE model revealed the worst results.

# Iterative Motif Scanning in HMM Framework

**Maja Zagorščak** (Croatia)
e-mail: maja.zagorscak@nib.si

Motifs are short protein substrings that are characterized by a specific mutation pattern. They are sometimes considered as basic units of molecular evolution, so their detection and description are of great importance. In this note, we present an iterative motif scanning method, that is, a procedure that detects all variations of a given string in a large set of protein sequences. The method combines two well-known motif-scanning procedures - Position Specific Scoring Matrix (PSSM) and Hidden Markov Model (HMM)-based posterior decoding - for increased accuracy and is based on a statistical analysis of each iteration.

With a sample (motif/s) and data set (proteome or similar) given, iterative procedure consists of the following steps:
i) building an initial motif profile using a model of amino acid evolution,
ii) scanning the data for the best profile match in each sequence,
iii) obtaining a list of positive hits for each iteration and
iv) building a new motif profile for a new scan using the list of positives until no further changes in the list could be detected.

It is shown that a PSSM score can be described as a difference of two Gumbel distributed random variables, so PSSM scores will be approximately logistically distributed and positive hits will be those within the heavy tail of the distribution. Also, almost identical scan results, obtained by PSSM and HMM posterior decoding, are required per each iteration.

The method is tested on several well-known data sets.

# Participants

Ahlin Črt (Slovenia): crt.ahlin@gmail.com
Azzolina Danila (Italy): danila.83@live.com
Balogh Kitti (Hungary): kttblgh@gmail.com
Batagelj Vladimir (Slovenia): vladimir.batagelj@fmf.uni-lj.si
Belak Tamara (Slovenia): Tamara.Belak@gov.si
Benke János (Hungary): jbenke@math.u-szeged.hu
Berchialla Paola (Italy): paola.berchialla@unito.it
Bernard Sonja (Slovenia): bernard.sonja@gmail.com
Blejec Andrej (Slovenia): Andrej.Blejec@nib.si
Bolyog Beáta (Hungary): bbeata@math.u-szeged.hu
Bősze Zsuzsanna (Hungary): b.zsuzsi94@gmail.com
Burger-Ringer Luzia (Austria): l.buri@gmx.at
Cimmino Rita (Italy): ritacimmino@gmail.com
Cugmas Marjan (Slovenia): marjan.cugmas@fdv.uni-lj.si
Čuk Jerneja (Slovenia): Jerneja.Cuk@gov.si
Dumičić Ksenija (Croatia): kdumicic@efzg.hr
Fazekas István (Hungary): fazekas.istvan@inf.unideb.hu
Ferligoj Anuška (Slovenia): anuska.ferligoj@fdv.uni-lj.si
Friedl Herwig (Austria): hfriedl@tugraz.ac.at
Gür Sercan (Austria): sguer@wu.ac.at
Hermann Philipp (Austria): Philipp.Hermann@jku.at
Holter Magdalena (Austria): magdalena.holter@medunigraz.at
Jauck Stefanie (Austria): stefanie.jauk@medunigraz.at
Karasek Sarah (Austria): s.karasek@tugraz.at
Károlyi Kristóf (Hungary): karolyi.kristof@gmail.com
Kolak Anja (Slovenia): anja.kolak@fdv.uni-lj.si
Košmrlj Katarina (Slovenia): katarina.kosmrlj@gmail.com
Kovačić Jelena (Croatia): jkovacic@imi.hr
Kuchling Sabrina (Austria): sabrina.kuchling@ages.at
Kušar Maša (Slovenia): mk6221@student.uni-lj.si
Lalovič Marko (Slovenia): marko.lalovic@yahoo.com
Lanera Corrado (Italy): corrado.lanera@unipd.it
Lukácsné Porvázsnyik Bettina (Hungary): porvazsnyik.bettina@inf.unideb.hu
Lusa Lara (Slovenia): lara.lusa@mf.uni-lj.si
Manevski Damjan (Slovenia): dm5439@student.uni-lj.si
Matjašič Miha (Slovenia): mat.miha@gmail.com
Németh Renáta (Hungary): nemethr@tatk.elte.hu
Okorn Rok (Slovenia): okorn.rok@gmail.com

Omerovic Sanela (Austria): s.omerovic@student.tugraz.at
Omladič Matjaž (Slovenia): matjaz@omladic.net
Pavlič Klemen (Slovenia): klemen.pavlic@mf.uni-lj.si
Pecsora Sándor (Hungary): pecsora.sandor@inf.unideb.hu
Perecsényi Attila (Hungary): perecsenyi.attila@inf.unideb.hu;
Pikelj Jerneja (Slovenia): Jerneja.Pikelj@gov.si
Pivar Jasmina (Croatia): jpivar@efzg.hr
Pršlja Antonija (Slovenia): antonija.prslja@gmail.com
Pršlja Katarina (Slovenia): prslja.katarina@gmail.com
Pršlja Paulina (Slovenia): paula.prslja@gmail.com
Rakovics Márton (Hungary): rakovicsmarci@gmail.com
Rudas Tamás (Hungary): rudas@tarki.hu
Rustja Helena (Slovenia): helena.rustja@gmail.com
Ružić Nina (Slovenia): nina.ruzic@mf.uni-lj.si
Slavec Ana (Slovenia): ana.slavec@fdv.uni-lj.si
Srakar Andrej (Slovenia): andrej_srakar@t-2.net
Stare Janez (Slovenia): janez.stare@mf.uni-lj.si
Szák-Kocsis Csilla (Hungary): szakcsilla@gmail.com
Toman Aleš (Slovenia): ales.toman@ef.uni-lj.si
Wallisch Christine (Austria): christine.wallisch@meduniwien.ac.at
Zagorščak Maja (Croatia): maja.zagorscak@nib.si
Žnidaršič Anja (Slovenia): Anja.Znidarsic@fov.uni-mb.si