NBBC23 Abstracts

Key Note Presentations

Personalized prediction of drug response in cancer

Yudi Pawitan
Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Sweden.

The goal of personalized or precision medicine is to provide a treatment – including type of drug and dosing -- that is optimized for each patient according to their personal characteristics. Overall, this is hard to achieve, so across many diseases, most patients end up getting the same treatment. Cancer, however, is different. In addition to tissue specificity, the current molecular techniques can often reveal the specific oncogenic mutation(s) that drive cancer in a specific individual. For instance, almost all chronic myeloid leukemia (CML) is due to the oncogenic BCR-ABL fusion. Theoretically, if (a big if) a cancer depends on a specific driver gene or pathway, a therapy that targets the driver will kill the cancer. Much of the current effort in cancer therapy is focused on developing target-specific inhibitors. Given the wealth of genomic data we can collect on each tissue sample and a wide selection of inhibitors, personalized cancer therapy is becoming increasingly feasible. However, despite some encouraging successes, predicting the therapy response remains highly challenging due to underlying tumor heterogeneity, including natural evolution and potentially multiple drivers.

I will describe our efforts in the prediction of drug response of acute myeloid leukemia (AML). It's an ideal model to study because of its high molecular heterogeneity, and there is a rich collection of molecular and drug response data from the BeatAML consortium (n=461 patients, 122 drugs). We obtained a validated Spearman correlation of 0.68 (95% CI: 0.64-0.68) between the observed and predicted responses. Among predictions with a confidence score >0.75, the validated proportion of good responders is 77%. I will also describe a continuing investigation of combination therapies, where we try to identify multiple drugs that have synergistic interactions in their effects, potentially enlarging the search space for effective therapies.

Issues in data-monitoring for complex clinical trials

Deborah Ashby
Imperial College London, UK.

During a clinical trial, accruing data is often seen in confidence by a data monitoring committee to evaluate whether early termination of the study or other modifications are needed in the light of emerging results. For a classical two-group trial with a single primary endpoint there are well-established statistical approaches. However, more complex trials, such as platform trials or trials with multiple endpoints, present different challenges. Such designs are now being used more regularly in practice. We review these designs and current approaches, as well as challenges that are emerging, and outline where further developments are needed.

Covid-pandemic as a lesson of scientific communication (for statisticians) and statistical literacy (for politicians, media and everyone else): the Estonian experience.

Krista Fischer
[1] Institute of Mathematics and Statistics, University of Tartu, Estonia.
[2] Estonian Genome Center, Institute of Genomics, University of Tartu, Estonia.

The emergence of SARS-CoV2 virus and Covid-19 pandemic in spring 2020 affected the work and private lives for virtually everyone. For many statisticians, regardless of their previous experience with infectious disease data, it brought unprecedented data analytic and scientific communication challenges. There was an urgent need to provide evidence-based support for policy-makers at national government levels, whereas the data and even underlying assumptions were constantly changing. As an additional challenge, the scientific experts became also pressurized by the media, with constant requests to provide information on new data and prediction results. At the same time the statistical literacy level of policy-makers, media and the general public was often not sufficient for efficient 2-way communication. Now, three years later we can summarize the experience and hopefully be better prepared to provide statistical support for national governments in a crisis situation next time (if needed). Also, there is still a lot of data available and interesting research questions to be answered on the pandemic.

In this talk, an overview of the Estonian experience will be given, from a viewpoint of a statistician (and a member of the Estonian Covid-19 Scientific Advisory Board, formed by the government of Estonia).

Invited Sessions

Precision medicine

Borrowing Information — Between Patients, Subpopulations and Trials.

Carl-Fredrik Burman, AstraZeneca

In a dream scenario, every patient receives the treatment that is best for her, depending on measurable covariates as well as on her preferences. In practise, the best individualisation of treatments is often unknown. When trying to predict treatment effects for a certain patient, we generally have to generalise from responses in other patients. Prediction modelling thus requires borrowing information from a wider population. The concept of information borrowing is relevant in many related situations. Several randomised clinical trials (RCTs) have pre-specified subpopulations, and may lead either to marketing authorisation in the full population or e.g. in biomarker positive patients only. Often the biomarker negative subpopulation is too small to have reasonable power in its own right, and borrowing between subpopulations may be a solution. Borrowing can also occur from 2-arm trial data in a wide population to a smaller population (e.g. adults to children), often using historical trials. Alternatively, historical control group data can be borrowed to complement the control group in an RCT, e.g. in cases of (ultra-)rare diseases where trial sizes are limited. Digital Twin methodology may use machine learning to train a predictive model on historical data for patients on standard-of-care treatment, to improve precision in a prospective RCT. In this presentation, we discuss information borrowing in general and give examples from different contexts. We will stress issues around fundamental concepts in statistical inference, as the type 1 error and its definition, the role of assumptions, and Bayesian vs. frequentist thinking.

Gene-environment interactions in metabolic traits: challenges and opportunities for personalised medicine

Anders Eriksson, Tartu Univ.

Metabolic traits play key roles in health. For example, overweight and obesity are major risk factors for a number of chronic diseases. However, despite a well-known and significant contribution of genetic factors in obesity, current prediction tools based on individual genetic data (polygenic scores) add only limited value to existing methods based on lifestyle and other clinical risk factors in predicting obesity. Recent work has demonstrated that effects of genetic variants on individual's obesity risk is substantially affected by lifestyle factors, such as level of physical activity, diet, and the interaction between those factors, but such gene-environment interactions are typically not considered in classical quantitative genetics approaches, resulting in reduced prediction accuracy and transferability across cohorts. Statistical frameworks that use machine learning approaches to combine rich environmental data with genetic approaches offers a way forward and has the potential to transform our ability to identify patients most at risk and the most effective prevention and treatment strategies for the individual.

Enhancing Randomized Clinical Trial publications to simplify population adjustment and personalized medicine

Mario Ouwens, AstraZeneca

In Health Technology Assessment, Randomized Clinical Trials (RCT) from different treatments are combined in order to indirectly compare treatments which are not compared in one trial. This indirect treatment comparison is not a randomized comparison and estimated effects may therefore partly be driven by the different treatments, but also by differences in distribution of effect modifiers. Usually, one has access to individual patient data from own trials, but not from comparator trials. To account for potential effect modification, while only having aggregated data, methodology is developed, among which Matching Adjusted Indirect Comparisons, Simulated Treatment Comparisons, ML-NMR and Network Meta Imputation. These methods may not have been needed when RCT publications would provide functional relationships assigning treatment effect to patient characteristics. If RCT publications would include these, the functional relations can be filled in for new trials, to get an estimated treatment effect per person in the new trial. When all trials would provide these relationships, one can perform network meta analyses from which treatment prescription could be optimized. This would help health care professionals and ultimately patients as well. There is quite a reluctance against this type of methodology, because RCTs are not powered for subgroup analyses and rules to assess whether a patient characteristic is an effect modifier are not widely known/non-existing. However, first steps in this direction are made by having forest plots with subgroup results, potentially leading to licensing for subgroups. This presentation will promote the idea of including the functional relationships in RCT publications, the impact it can have on HTA submissions and personalized medicine as well as pro's and con's of this idea.

Causal Inference

A nonparametric test for treatment effect for survival data when censoring may depend on treatment as well as covariates.

Torben Martinussen, University of Copenhagen

We consider the situation with a survival endpoint and is interested in testing if there is evidence of a treatment effect in the setting where censoring may depend on both the treatment and covariates. It is well known that traditionally test, such as the log-rank test, breaks down if censoring is not completely independent. Tests based on semiparametrically modeling work if either the proposed model is correctly specified or, and maybe not so well known, if the treatment is randomized. We start by revisiting the latter result and then move on to the general setting where we do not assume any specific model nor require that treatment is randomized. The proposed method builds on the newly suggested assumption-lean inference by Vansteelandt and co-workers (JASA, 2022).

Continuous-time TMLE for causal inference in time-to-event settings

Helene Charlotte Wiese Rytgaard, University of Copenhagen

Targeted learning (TMLE) provides a general framework for estimation of causal parameters, combining machine learning estimation of high-dimensional nuisance parameters with rigorous statistical inference obtained via a targeted update step. The continuous-time TMLE is a generalization of the targeted learning methodology for nonparametric causal inference in settings where interventions, covariates and outcome change at subject-specific points in time. In this talk, I will discuss the overall ideas of the continuous-time TMLE framework, highlighting both its potential and challenges. As a specific example, I will focus on its use for the estimation of causal effects of time-fixed treatment decisions on absolute risk probabilities in classical time-to-event settings.

Indirect comparison with proximal causal inference

Zehao Su, University of Copenhagen

Indirect comparison of treatments is an important instance of evidence synthesis for health technology assessment. Yet there are two major challenges for existing indirect comparison methods: the estimand may not have an easy, possibly causal interpretation, and the presence of unobserved prognostic variables or effect modifiers undermines the validity of transportability assumptions. Building upon the recent proximal causal inference framework (Miao, Geng, et al., 2018), we present nonparametric identifiability conditions and estimation methods for a causally-interpretable estimand under the presence of unobserved prognostic variables. When appropriate negative controls are available in trial data, the treatment-trial and/or the outcome bridge functions, which identify the target parameter by accounting for the lack of knowledge of unobserved variables, can be estimated from the observed data. We show that the efficient-influence-function-based estimator is doubly robust against the misspecification of bridge functions. A numerical study illustrates the behaviour of the estimator under finite samples. We also provide a cautionary note for the use of proximal indirect comparison when unobserved effect modifiers are of interest.

Modelling based on finite mixtures

Modelling wage earnings preceding disability retirement using trajectory analysis

Janne Salonen, Finnish Public Sector Pension Provider Keva, Finland

The question of work-ability preceding disability retirement is essential in labor market policy in preventing permanent disability pensions. Labor market attachment measured by wage earnings (i.e. employment) is one predictive indicator of future work-ability. We aim to identify the high or low risk groups over life-course. This study focuses on wage earnings trajectories before disability retirement, that is of those who will retire.

The research data is collected from the administrative registers of Finnish public sector pension provider Keva, and it consists of a random sample of local government disability retirees in 2017 (n=448). The study-design consist of a follow-up of wage earnings over years 2005–2017. Trajectory analysis (e.g. Nagin 2005) is used to identify sub-population of retirees using wage earnings as a single outcome. The overall distribution of wage earnings is highly right-skewed with high share of zeros, which both challenge the statistical analysis. The central task for the statistical analysis is to identify the correct number sub-populations with different earnings profiles. We apply the scaled transformation (Spitzer 1984; Gurka et al. 2006) which is closely related to Box-Cox transformation (Box and Cox 1964) for the wage earnings to determine which transformation is implied by the data. The transformation parameters and statistical information criteria (e.g. BIC) indicate a log transformation for the wages. In practice the analysis with the non-transformed wages yields a trajectory analysis solution of eight groups, whereas the transformed wage yields six trajectory groups.

The substantial key results can be summarized by stating that there are three large groups whose wages are not affected by the upcoming disability retirement. These groups constitute the main part of the sample (69 per cent). For two groups the wages are strongly affected by the forthcoming permanent disability (total 22 per cent). For the last group (9 per cent), where wages have initially risen strongly over study-period, the earnings slightly fall before permanent disability. Group-specific background information (e.g. cause of disability) is also given.

On the improved estimation of normal mixture components for longitudinal data

Tapio Nummi, Tampere University/ITC faculty, Finland; Jyrki Möttönen, Department of Mathematics and Statistics/Helsinki University, Finland; Janne Salonen, Finnish Public Sector Pension Provider Keva, Finland; Pasi Väkeväinen, Tampere University/ITC faculty, Finland; Timothy E. O'Brien, Department of Mathematics and Statistics, and Institute of Environmental Sustainability, USA

In many modeling applications a finite normal mixture (see e.g. Everitt and Hand, 1981) as well as its extensions, such as mixture skew-normal distribution (Lin et al. 2007b) or mixture t-distribution (Lin et al. 2007a), provides a sensible model for the data at hand. However, finding the best possible component distribution among many competing alternatives can be a very complicated and challenging task. For example, the number of components needed may depend on the sample size, assumed component distribution as well as the possible transformation (scale of measurements) applied. Actually, two slightly different goals can be identified: Approximation of the distribution of the response variable and identification of the number of sub-populations.

The focus of this talk is on so-called trajectory analysis (TA) that is an application of finite mixture modeling for longitudinal data (Nagin, 1999 and 2005). We propose a method that is based on the scaled Box-Cox transformation (Spitzer, 1984 and Gurka et al. 2006) that makes the likelihood based inference possible over transformed responses and the actual analyses are then based on the components of transformed normal mixtures. The proposed approach has several advantages. First, the theory of TA with normal mixtures is well established and many implementations as software packages already exists. Second, a suitable transformation can reduce the risk of generating the excess number of mixture groups. The data analytic part of the talk is based on real data applications of birth weight, trajectories of alcohol consumption as well as simulation experiments.

References:

Everitt, B. and Hand, D. (1981). Finite Mixture Distributions. London: Chapman and Hall.

Gurka, M., Edwards, L., Muller, K. and Kupper, L. (2006). Extending the Box-Cox transformation to the linear mixed model. Journal of the Royal Statistical Society, A (Statistics in Society), Volume 169, Issue 2, p. 273-288.

Lin, T., Lee, J. and Hsieh, W. J. (2007a). Robust mixture modeling using the skew t-distribution. Statist. Comput., 17, p. 81-92.

Lin, T., Lee, J. and Yen, S. (2007b). Finite mixture modeling using the skew normal distribution. Statistica Sinica, 17, p. 909-927.

Nagin, D. (1999). Analyzing developmental trajectories: Semi-parametric, group-based approach. Psychological Methods, 4, p. 139-177.

Nagin, D. (2005). Group-based modeling of development. Cambridge, MA: Harvard University Press.

Spitzer, J. (1984). Variance estimates in models with the Box- Cox transformation: Implications for estimation and hypothesis testing. Review of Economics and Statistics, 66. p. 645-652.

Variable selection in mixture regression for longitudinal data based on joint mean-covariance model

Jianxin Pan, BNU-HKBU United International College, China

A large number of explanatory variables may be measured with the collection of longitudinal data, of which some may not be influential for modeling of heterogeneous longitudinal data. For such complex data, not only their mean but also covariances may be affected by various explanatory variables. A data-driven approach is proposed to model the mean and covariance structures, simultaneously, together with selecting influential explanatory variables. A penalized maximum likelihood method for the joint mean and covariance model is developed within the framework of finite Gaussian mixture regression. EM algorithm is employed for its numerical calculation. The parameter estimators obtained are shown to be consistent and asymptotically normally distributed, and have oracle properties with proper choices of penalty function and tuning parameter. Simulation studies show that the proposed method works very well and provides accurate and effective parameter estimators by conducting variable selection. For illustration, real data analysis on clustering COVID-19 infected cases for European countries in terms of governmental

Teaching and consulting

The power of a joint effort – consultancy through a network of biostatisticians, bioinformaticians and data analysts

Annica Dominicus, Karolinska Institutet, Sweden

With an analytic and curious mind, and some knowledge of study design and statistical methodology, it is often possible to help medical researchers to refine the research question, and guide the project in the right direction through advisory meetings in an early stage of the project. Of course, with knowledge about the specific medical field and a deeper understanding of specific study designs and methodologies that may be relevant to the project, even better advise can be provided. How is it possible to provide such high quality consultancy to clinical and translational researchers connected to a large medical university and health care providers in a whole city region with limited resources? Our solution is to make it a joint effort through a network of experts. We describe the ongoing implementation of CLINICUM in the Stockholm region and the challenges and learnings so far.

Understanding transition from gestational diabetes to Type-2 diabetes with ML

Roza Maghsood, AstraZeneca, Sweden

PONCH (Pregnancy Obesity Nutrition and Child Health) project is a collaboration between AZ and Gothenburg university to explore and analysis the data available from women with gestational diabetes (GDM) diagnosis. This population (women with GDM) have a high risk of developing T2D within a relatively short timeframe and the idea is to learn more about factors that drive development of T2D in the general population. We have a retro dataset available with 6 and 10 years post gestational diabetes (GDM) diagnosis and there will be a new dataset in the future. The idea is to use some ML approaches for the current dataset to see if the pre-defined classes like NGT, IGT, T2D and T1D are distinguishable. The approach can be useful to find the main features like BMI, glucose, insulin, Hba1c and HOMA-IR in transition from gestational diabetes to Type-2-diabetes.

Setting up a new MSc programme in Biostatistics and Data Science within Stockholm Trio

Therese M-L Andersson, Karolinska Institutet, Sweden

To secure the future growth of biostatistics within the Nordic-Baltic region, it is important to ensure that there exist education programs to attract new individuals to the field. At the moment, there are no master’s programs in biostatistics in Sweden, however it was recently decided that a new program in biostatistics and data science will start in 2024. The program will be jointly run by three universities in Stockholm, Karolinska Institutet, Stockholm University and KTH Royal Institute of Technology, commonly referred to as Stockholm Trio. The students will receive a rigorous training in statistics, computational science, and programming along with the theoretical and practical education to enable them to apply these skills to address challenges that arise in biology, medicine, and public health. This will be done by taking advantage of the strengths from each of the participating universities. In this presentation I will describe the plans for the program, and the ongoing implementations.

Combination of (Very) Different Data Sources

Independent Inspection, Confidence Conversion, Focused Fusion: combining (very) different information sources

Nils Lid Hjort, Department of Mathematics, University of Oslo

I present a general framework for combining information across potentially very different information sources, required to handle more difficult setups than the typical meta-analysis situations. I then briefly illustrate this general II-CC-FF methodology of Cunen and Hjort (Scandinavian Journal of Statistics, 2021) for two biostatistical applications.

Blending forecasts for the time-to-frost

Céline Cunen, Norwegian Computing Centre, Oslo

In this talk I will present some general methods for combining probabilistic forecasts from two or more sources. Ideally, the final blended forecast should be well-calibrated and more precise than any of the single-source forecasts. We have developed a new blending method, called the Gaussian Forecast Filter (GFF), which seems to have good properties. The methods will be illustrated through a real application: attempting to precisely forecast the time-to-frost some months ahead by blending seasonal and subseasonal forecasts. This talk is based on joint work with Thea Roksvåg, Claudio Heinrich-Mertshing and Alex Lenkoski.

Interpretable multi-omics integration using sparse joint matrix factorization

Felix Held, Department of Mathematical Sciences, Chalmers University of Technology and University of Göteborg

Interest in unsupervised methods for joint analysis of heterogeneous data sources has risen in recent years. This is in part due to the increased availability of multi-omics data collected on multiple groups of individuals (patients, cells). Low-rank latent factor models for multi-view (one group, multiple data types) and grid-like (multiple groups, multiple data types) layouts have been shown to be effective for this task. Of particular interest are the separation of the variability between data sources into joint, partially shared, and individual components, as well as the interpretability of estimated latent factors.

We present a novel computational method for the estimation of low-rank latent factor models that supports missing values, flexible layouts, separation of variability, as well as the estimation of interpretable factors. Flexible layouts encompass traditional multi-view and grid-like layouts and in addition allow relationships between groups/data sources or grid-like layouts with missing blocks. Using a penalized approach, we automatically separate variability into the desired components and estimate sparse representations of factors. To increase the interpretability of factors, we enforce orthogonality. Estimation is performed efficiently in an ADMM-type optimization framework which we adapted to support (some) manifold constraints. The performance of our method is demonstrated on simulated data and examples of a real data analysis are shown.

Contributed sessions

Biostatistical modelling

Bayesian gene regulatory network inference using mixture priors

Emre Karaman, Center for Quantitative Genetics and Genomics, Faculty of Technical Sciences, Aarhus University, DK-8000 Aarhus C, Denmark

Revealing complex regulatory relationships among genes and the mechanism behind gene expression is important in many biological context, such as agriculture and medicine. Several computational methods exist for inferring genes' regulatory networks (GRNs) and finding expression quantitative loci, from gene expression and single-nucleotide polymorphism data. A limiting factor for the accuracy in GRN inference is that gene expression data are generally available only for a very small fraction of a population, while the number of parameters are much larger, n << p. In order to obtain unique estimates of model parameters, and enforce sparsity in the parameter matrix, Bayesian methods with mixture priors can be implemented. In this study, we will use simulations to test the accuracy of Bayesian GRN inference using a set of mixture priors. A state-of-the-art GRN inference method based on sparse structural equation models will also be used for comparison. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815668.

Sensitivity Analysis of G-estimators to Invalid Instrumental Variables

Valentin Vancak & Arvid Sjölander, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet

Instrumental variables regression is a tool that is commonly used in the analysis of observational data. The instrumental variables are used to make causal inference about the effect of a certain exposure in the presence of unmeasured confounders. A valid instrumental variable is a variable that is associated with the exposure, affects the outcome only through the exposure (exclusion), and is not confounded with the outcome (exogeneity). Unlike the first assumption, the other two are generally untestable and rely on subject-matter knowledge. Therefore, a sensitivity analysis is desirable to assess the impact of assumptions' violation on the estimated parameters. In this paper, we propose and demonstrate a new method of sensitivity analysis for G-estimators in causal linear and non-linear models. We introduce two novel aspects of sensitivity analysis in instrumental variables studies. The first is a single sensitivity parameter that captures violations of exclusion and exogeneity assumptions. The second is an application of the method to non-linear models. The introduced framework is theoretically justified and is illustrated via a simulation study. Finally, we illustrate the method by application to real-world data and provide guidelines on conducting sensitivity analysis.

A FRAMEWORK FOR COUNTERFACTUALS IN MARKEDPOINT PROCESS MODELS

Pål Christie Ryalen

Many researchers use counterfactual variables for causal reasoning. Frameworks such as the FFRCISTG model formalizes the study of a finite number of such variables. A corresponding framework of counterfactuals in marked point process (MPP) models, which deals with events that occur in continuous time, has not yet been developed. This is of interest in fields such as survival and event history analysis, where 1) there is a large literature of statistical methods formulated in continuous time, and 2) interest is frequently in effects of time-varying treatment strategies, which may often be most appropriately viewed as time-continuous processes.

In this presentation, we aim to fill this gap by providing a rigorous framework for counterfactuals in MPP models. Specifically, we introduce *strong causal realizations*, which serve as counterfactual processes. We provide analogs of the existing ’exchangeability’ and ’consistency’ assumptions among these processes. We also establish nonparametric identification results in the presence of unmeasured confounders, including a continuous-time g-formula and IPW formula. We additionally describe and discuss time-varying treatment regimes in this setting. Finally, we show how the method can be used to reason about identification of commonly studied parameters in survival analysis, and highlight some similarities and differences with existing methods.

Causal graphs for identification of causal effects in continuous-time event-history analyses

Kjetil Røysland, Vanessa Didelez and Pål C Ryalen

We consider continuous-time survival or more general event-history settings, where the aim is to infer the causal effect of a time-dependent treatment process. This is formalised as the effect on the outcome event of a (possibly hypothetical) intervention on the intensity of the treatment process, i.e. a stochastic intervention. To establish whether valid inference about the interventional situation can be drawn from typical observational, i.e. non-experimental, data we propose graphical rules indicating whether the observed information is sufficient to identify the desired causal effect by suitable re-weighting. In analogy to the well-known causal directed acyclic graphs, the corresponding dynamic graphs combine causal semantics with local independence models for multivariate counting processes. Importantly, we highlight that causal inference from censored data requires structural assumptions on the censoring process beyond the usual independent censoring assumption, which can be represented and verified graphically. Our results establish general non-parametric identifiability and do not rely on particular survival models.

Analysis of pension and disorder data

Performance evaluation of machine learning techniques for prediction of disability pension among workers with Musculoskeletal Disorders

Adnan Noor Baloch, Helena Sandén, Mats Hagberg, Martin Adiels

Purpose

Disability pension due to musculoskeletal disorders is a public health problem. Beside musculoskeletal symptoms, work environment and demographical factors are important contributors to disability pension. Although machine learning techniques are becoming increasingly popular in predictive modelling and clinical decision-making, we do not know about the relative performance of these techniques. In this study, we examine the performance of seven machine-learning techniques for predicting disability pension among workers with musculoskeletal disorder in Swedish workforce.

Methods

Swedish Work Environment survey(2009 – 2011) was linked to Swedish official register on disability pension. A total of 9370 workers reported musculoskeletal symptoms and 377 were on disability pension during the follow-up year. Two models (Model-1 with 7 and Model-2 with 14 predictors) were built. The performance of multivariable logistic regression was compared with Random forest, AdaBoost, Gradient boosting machine, Extreme gradient boosting, Support vector machine and Naïve Bayes classifier for each model. The hyperparameters of machine learning techniques were tuned. The data was randomly split in two pieces with 65% for the training set and 35% for the validation set. The discrimination (area under ROC) and calibration (Spiegelhalter z-test, Brier score) of the validation set were used as the performance metrics.

Results

All techniques performed excellently on the training set(area under ROC, 0.84 — 0.90) with a decrease albeit satisfactory performance on the validation set(area under ROC, 0.81 — 0.87). Adding demographical data(Model-2) increased the discrimination ability (area under ROC) by an average of 3% compared with only work-environment data(Model-1). All techniques demonstrated good calibration (Brier score<0.04, non-significant Spiegelharter z-test).

Conclusions

Our results indicate that logistic regression with clinically relevant predictors meets the performance of advanced machine learning techniques. The studied techniques had similar performance. Machine learning techniques can be employed to estimate personalized probabilistic predictions. Adding demographic data improved the performance of all techniques.

Sickness absences and earnings before the disability pension application: A trajectory analysis of Finnish municipal sector employees

Petra Sohlman, Finnish Public Sector Pension Provider Keva, Finland

Regarding the Finnish national objective of extending working lives and preventing early exits from labor market, the years leading to a disability pension are an important time frame for preventive action. It is essential to recognize those employee groups which are in the greatest risk of permanently losing their work ability. The aim of this study was to describe the differences in sickness absence patterns among municipal sector employees by examining the latent trajectories in annual cumulative sickness absence days and wage earnings. Consequently, the locations of those employees filing an application for disability pension in these trajectory groups was examined, with focus on the underlying sickness or disability.

The data were gathered from the administrative registers of the Finnish public sector pensions provider Keva and municipal employers. The data consist of the employment spells, sickness spells, occupations, and wages of 113 410 municipal sector employees during the period of 2016-2021. A follow-up of first-time disability pension applications filed during 2021-2022 was made to indicate the underlying illness, the type of pension applied for, and the decision of the pensions provider. The data on sickness spells used here include all sickness-related absences from work, also short spells which usually fall outside the scope of research data. Also in this study, the application for a disability pension is used as a follow-up instead of the passage into a pension, which allows for also including those employees who end up receiving a rejecting decision but still carry a doctor's evaluation of diminished working ability.

Trajectory analysis identified a total of eight subgroups from the population using annual earnings and annual cumulative sickness days as outcomes. Two outcomes were used to make observations on the development on earnings as the amount of annual sickness absence varies. The BIC value guided the selection of the number of subgroups. These subgroups differ in terms of job title, age group, and level of earnings. Those ending up with applying for a first-time disability pension belonged mostly to the increasing absences or persistent high absences trajectory groups, with no effect on annual earnings during the years leading to the application. Those who received a rejecting decision belonged mostly to the persistent high absences group. There are some differences between the diagnostic classes in the distribution into subgroups. Those with musculoskeletal diseases more often than those with mental disorders follow the persistent high absences trajectory. Those with diseases of the nervous system follow the trajectory of continuous low annual sickness absences in fifth of the cases, while those with tumors follow the increasing absences trajectory in half of the cases.

On the Performance Comparison of Ordinal Regression Models

Deniz Sigirli[1], Barbora Kessel[2], Anna Grimby-Ekman[2]
1 Bursa Uludag University, Faculty of Medicine, Department of Biostatistics, Bursa, Turkey
2 School of Public Health and Community Medicine, Institute of Medicine, Gothenburg University, Gothenburg, Sweden

Ordinal scales are common in for example public health and medical sciences, for assessing health symptoms, diagnostic ratings based on different radiologic modalities, stages of diseases, severity of pain, etc. Ordinal regression models allow to model the dependence of an ordinal response variable on a set of predictors, thus taking into account the ordinal structure of the response (1).

There are several ordinal regression models that can be used for the analysis of ordinal outcome data, including cumulative link models, continuation-ratio models, adjacent-category models or ordered stereotype models. While these different methods can be used to estimate and interpret the probabilities related to outcome variable in different ways, the most commonly used ones are the cumulative link models.

In cumulative link models an underlying latent variable structure is sometimes assumed, and a link function is being used to represent the relationship between the linear predictor and cumulative probabilities. However, there are some contradictions in the literature regarding choosing an appropriate link function in ordinal regression models (2-4). While comparing cumulative link models constructed with different link functions, it would not be proper to use goodness-of-fit tests or likelihood and deviance-based test statistics since these can be used to compare nested models. Receiver operating characteristic (ROC) analysis is a standard procedure for assessing the performance of binary classifier's performance and area under the ROC curve (AUC) have been extensively used as a performance metric as a single scalar measure. Recently, different approaches have been proposed to extend ROC analysis for multi-class classification and estimate volume under the ROC surface (VUS) for more than two classes (5-8). While in binary classification models, the predicted probability outcome can be used when estimating AUC, there is no such natural ordering structure of the predicted values of ordinal regression models. Waegeman et al. (2008) proposed using $\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X}$ values obtained from cumulative link models as a rank function that ensures objects with higher $\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X}$ values are classified to higher outcome category.

We illustrate the use of the estimated VUS, in comparing the choice between different link models, in an applied example based on data from a pain study. We also use simulations to examine the capability of VUS to support the choice of the link function.

References
1. Agresti, A. (2010) Analysis of Ordinal Categorical Data. 2nd Edition, Wiley, Hoboken.
2. Norusis, J. M. (2012). IBM SPSS statistics 19.0 advanced statistical procedures companion. Upper Saddle River, NJ: Prentice Hall.

3. Li, K.C. & Duan, N. "Regression Analysis Under Link Violation." Ann. Statist. 17 (3) September, 1989.

4. Thomas, S.J., Walker, D. & C.M., McKenna. (2020). An Exploration of Link Functions Used in Ordinal Regression. Journal of Modern Applied Statistical Methods. 18. 2-15.

5. Mossman D. Three-way ROCs. Med Decis Making. 1999 Jan-Mar;19(1):78-89.

6. Dreiseitl S, Ohno-Machado L, Binder M. Comparing Three-class Diagnostic Tests by Three-way ROC Analysis. Medical Decision Making. 2000;20(3):323-331.

7. Nakas CT & Yiannoutsos CT. Ordered multiple-class ROC analysis with continuous measurements. Stat Med. 2004 Nov 30;23(22):3437-49. doi: 10.1002/sim.1917.

8. Xiong C, van Belle G, Miller JP, Morris JC. Measuring and estimating diagnostic accuracy when there are three ordinal diagnostic groups. Stat Med. 2006 Apr 15;25(7):1251-73.

9. Waegeman, W., De Baets, B., & Boullart, L. (2008). ROC analysis in ordinal regression learning. Pattern Recognition Letters, 29(1), 1–9.

Health and epidemiology

PROJECTING THE HEALTH CARE COSTS AFTER OBSTRUCTIVE SLEEP APNEA DIAGNOSIS USING REGISTER DATA AND BAYESIAN PREDICTIVE MODELLING

Jukka Kontto[1], Reijo Sund[2], Tommi Härkänen[1]
1 Population Health Unit, Finnish Institute for Health and Welfare, Helsinki, Finland
2 Institute of Clinical Medicine, University of Eastern Finland, Kuopio, Finland

Background: Obstructive sleep apnea (OSA) is the most frequently encountered sleep-related breathing disorder. It has been estimated that OSA afflicts approximately 1 billion people worldwide, and the prevalence of OSA is increasing (Alakörkkö 2022). We aim to project the health care costs of OSA patients after their diagnosis in Finland. The work has been conducted within the Sleep Revolution project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 965417.

Data: We used register data containing Finnish patients diagnosed with OSA during 2016–2018 along with their outpatient health care costs from 2015–2020 and mortality information.
Methods: Our projection method exploits longitudinal data and Bayesian modelling with Markov chain Monte Carlo (MCMC) methods (Gelman et al. 2013). The health care costs of OSA patients are modelled using a hierarchical two-part logistic and gamma regression model. The posterior distribution of the model parameters is based on the observed repeated measures and time-to-event data. The individual-level projections are generated from the posterior predictive distribution using forward sampling. Total costs of all OSA patients, and by gender and age group are then calculated for every six months until five years after OSA diagnosis based on the individual predictions.

Results: Preliminary results show that the total costs of women remain lower compared to men. Also, the total costs of younger age groups remain lower compared to older age groups. We further evaluate our findings by comparing them to results generated using other methods.

References:
Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., & Rubin, D.B. (2013). Bayesian Data Analysis (3rd ed.). *Chapman and Hall/CRC*. https://doi.org/10.1201/b16018.
Alakörkkö I. (2022). The economic cost of obstructive sleep apnea – A systematic review (Master's thesis, University of Eastern Finland, Kuopio, Finland). http://urn.fi/urn:nbn:fi:uef-20220894.

A lean additive frailty model: with an application to clustering of melanoma in Norwegian families

Mari Brathovde[1,2], Tron A. Moger[3], Marit B. Veierød[2], Tom Grotmol[4], Odd O.
Aalen[2], Morten Valberg[1]
1 Oslo Centre for Biostatistics and Epidemiology, Oslo University Hospital, Oslo, Norway
2 Oslo Centre for Biostatistics and Epidemiology, Dept. of Biostatistics, Institute of Basic Medical Sciences, University
of Oslo, Oslo, Norway
3 Department of Health Management and Health Economics, University of Oslo, Oslo, Norway
4 Cancer Registry of Norway, Oslo, Norway

Large-scale health registries enable detailed studies of clustering of cancers in families. Frailty models provide a framework for conducting such studies at a much higher level of detail than conventional studies. However, these models' complexity grows with family size and the number of cancer diagnoses in each family. Consequently, these models have primarily been used in settings where cluster sizes are small, e.g. for twin pairs, or by considering only a few first-born children in a family. This poses a challenge for fully utilizing the detailed data available in the registries. We present a modification of the additive genetic gamma frailty model, which alleviates some of these problems by using a leaner additive decomposition of the frailty. The modified model is then used to analyze population-wide data on clustering of melanoma in 2,391,125 Norwegian families.

Using a first-order approximation of the genetic structure in nuclear families of parents and children, we obtain a model that reduces the complexity wrt. family size. Although a large number of cases within the same family still poses a challenge, this allows us to analyze a far greater class of datasets. An additional major benefit of the lean model is a significant speed-up in model fitting. This enables fitting even more complex models and makes model fitting on a desktop computer feasible without needing a high-performance cluster (HPC). We demonstrate in a simulation study that our proposed lean model gives a good approximation to the original additive genetic gamma frailty model.

Using the lean model, we can analyze the complete population-wide data set on melanomas in all Norwegian families registered from 1960-2016. We find a substantial clustering of melanomas in Norwegian families and a large heterogeneity in melanoma risk across the population. We estimated that there is a large inequality in frailty in the population, where 46% of the frailty could be attributed to the 10% of the population at the highest unobserved risk.

In conclusion, additive frailty models can be used to study relatively large clusters. Furthermore, there is a substantial clustering of melanomas in Norwegian families and a large heterogeneity in melanoma risk across the population.

Genetic associations vary across the spectrum of fasting serum insulin: results from the European IDEFICS/I.Family children's cohort

Kirsten Mehlig on behalf of the I.Family consortium
School of Public Health and Community Medicine, Institute of Medicine, University of Gothenburg, Gothenburg, Sweden

Objectives: Previous results from a Swedish cohort of middle-aged women showed a U-shaped relationship between fasting serum insulin and incident dementia (1). We aimed to perform genome-wide association (GWA) analyses for fasting serum insulin in European children with focus on variants associated with the tails of the insulin distribution.

Methods: Genotyping was successful in 2833 children from 7 European countries 2-14 years old at insulin measurement. Because levels vary during childhood analyses were based on pre-selected age- and sex-specific percentiles of fasting insulin modelled by logistic regression. Additive genetic models were adjusted for age, sex, BMI, year, country, and principal components to account for ethnic heterogeneity. We used quantile regression to examine how associations with specific variants varied across the insulin spectrum.

Results: GWA analyses of the tails of the insulin distribution identified several variants that were located on genes previously associated with both insulin and Alzheimer's disease (AD). One variant associated with the 85th insulin percentile was located on the *SLC28A1* gene previously linked to AD and type 2 diabetes (rs2122859, p-value = 3 10-8). Two variants associated with the 15th insulin percentile were located on *RBFOX1* (rs2109059, p = 9 10-7) and *SH3RF3* (rs36197836, p = 5 10-6), genes linked to brain amyloidosis and late-onset AD, respectively. Quantile regression showed large variability in effect size across the insulin spectrum for these variants.

Conclusions: Our approach identified variants that were associated with the tails of the insulin spectrum only. Because traditional heritability estimates assume that genetic effects are constant throughout the phenotype distribution, the new findings may have implications for the problem of missing heritability and the study of U-shaped biomarker-disease associations using Mendelian randomization.

References:

(1) Mehlig et al., Neurology 91(5):e427-e435 (2018)

Quantifying Movement Behavior of Chronic Low Back Pain Patients in Virtual Reality

TOMMI GRÖHN, Department of Computer Science, Aalto University, Finland; SAMMELI LIIKKANEN, DRDP, Institute of Biomedicine, University of Turku, Finland; TEPPO HUTTUNEN and MIKA MÄKINEN, EstiMates Oy, Finland; PASI LILJEBERG, Department of Information Technology, University of Turku, Finland; PEKKA MARTTINEN, Department of Computer Science, Aalto University, Finland

Chronic low back pain (CLBP) is a globally common musculoskeletal problem. Measuring the sensation of pain and the effect of a treatment has always been a challenge for healthcare. Here, we study how the movement data, collected while using a virtual reality (VR) program, could be used as an objective measurement in patients with CLBP. A specific data collection method based on VR was developed and used with CLBP patients and healthy volunteers. We demonstrate that the movement data in VR can be used to classify individuals in these two groups with a high accuracy by using logistic regression. The most discriminative features are the duration of the movements and the total variation of movement velocity. Furthermore, we show that hidden Markov models can divide movement data into meaningful segments, which creates possibilities for defining even more detailed features, with potential to improve accuracy, when larger datasets become available in the future.

Theory of statistics and biostatistics

Having a ball: Gaussian process regression as a probabilistic way to model and express spectator excitement of basketball matches

Andreas Kryger Jensen, University of Copenhagen

Many popular sports involve matches between two teams or players where each team score points throughout the match. While the overall match winner or result is interesting, it conveys little information about the underlying scoring trends throughout the match. Modeling approaches that accommodate a finer granularity of the score difference throughout the match are needed to evaluate in-game strategies, discuss scoring streaks, team strengths, and other aspects of the game. We propose a Bayesian Gaussian process model to express the score difference between two teams and introduce a Trend Direction Index as an easily interpretable probabilistic measure of the current trend in the match as well as a measure of post-game trend evaluation. In particular, we propose the Excitement Trend Index - the expected number of monotonicity changes in the running score difference - as a measure of overall game excitement. Our proposed method was applied to all 1143 matches from the 2019-2020 National Basketball Association season. We show how trends can be interpreted in individual games and how the excitement score can be used to cluster teams according to how exciting their games are to watch.

Large-deviation asymptotics of condition numbers of random matrices

Martin Singull, Denise Uwamariya, Xiangfeng Yang
Department of Mathematics, Linköping University, SE-581 83 Linköping, Sweden

Let $\boldsymbol{X}$ be a $p \times n$ random matrix whose entries are independent and identically distributed real random variables with zero mean and unit variance. We study the limiting behaviors of the 2-normal condition number $k(p, n)$ of $\boldsymbol{X}$ in terms of large deviations for large n, with p being fixed or $p = p(n) \rightarrow \infty$ with $p(n) = o(n)$. We propose two main ingredients: (i) to relate the large-deviation probabilities of $k(p, n)$ to those involving n independent and identically distributed random variables, which enables us to consider a quite general distribution of the entries (namely the sub-Gaussian distribution), and (ii) to control, for standard normal entries, the upper tail of $k(p, n)$ using the upper tails of ratios of two independent $\chi2$ random variables, which enables us to establish an application in statistical inference.

Edgeworth-type expansion of the density of the classifier when growth curves are classified via likelihood

Emelyne Umunoza Gasana[1], Dietrich von Rosen[2] and Martin Singull[1]
1 Linköping University, Linköping, Sweden
2 Swedish University of Agricultural Sciences, Uppsala, Sweden

When classifying repeated measurements using the Growth Curve model, also known as bilinear regression model, it can happen that the observations to classify might not belong to any of the two predetermined populations. [1] derived a two-step classification rule taking into account this perspective. Probabilities of misclassification of the two-step likelihood-based discriminant rule are established for the classification of growth curves where the distribution for the classifier is approximated using an Edgeworth-type expansion.

References:
[1] von Rosen, D. and Singull, M., Classification of repeated measurements using growth curves, Linköping University Electronic Press, 2022.