INTERNATIONAL BIOMETRIC SOCIETY

Invited Session Proposal

31ST International Biometric Conference

Riga, Latvia

We encourage the submission of proposals covering a wide range of topics in the theory and application of statistics to biological and life sciences. We look more favorably on sessions with a diversity of speakers with respect to gender, geographic region, and seniority. Session organizers may also be a speaker or discussant. Invited sessions are allocated 90 minutes.

Please submit your proposal for an Invited Session by email to the IBC2022 IPC Chair, Kerrie Mengersen at conference@biometricsociety.org.

Proposals should be submitted on or before 31 March 2021.

SESSION TITLE:
Prediction with observational data: STRATOS perspective

Key Words:
Predictive models; observational studies; model building; survival analysis; high-dimensional data; measurement error; initial data analysis; reproducibility; STRATOS initiative

PROPOSER/ORGANIZER:

Name: Michal Abrahamowicz and Willi Sauerbrei

Affiliations:
MA: Department of Epidemiology and Biostatistics, McGill University, Montreal, Canada;
WS: Institute of Medical Biometry and Statistic, University Medical Center Freiburg, Freiburg, Germany

Address:
MA: Research Institute of the McGill University Health Centre, Centre for Outcomes Research and Evaluation (CORE), 5252 boul. de Maisonneuve O, Montréal, Québec, Canada, H4A 3S5;
WS: Stefan-Meier-Straße 26, 79104 Freiburg, Germany

Email Addresses: michal.abrahamowicz@mcgill.ca; wfs@imbi.uni-freiburg.de

MOTIVATION:

Please describe the rationale for the topic of the proposed session.

Prediction is one of the most practically important but also most challenging tasks of statistical analyses. Indeed, the ultimate objective of many modern data analyses is to *predict* how, for example, the expected future health outcomes may be affected by alternative treatments, exposures and/or prognostic factors. Yet, to ensure the accuracy of the resulting predictions and the valid use of such predictions, researchers must address several complex analytical challenges, many of which require specialized statistical methods, often developed within different branches of statistical research. From this perspective, the proposed session aims at presenting and confronting separate views on some of the most practically relevant issues related to predictive modelling, by experts in *different* areas of modern biostatistics.

This approach is consistent with the overarching goal of the STRATOS (STRengthening Analytical Thinking for Observational Studies) Initiative, created in 2013 with the aim to systematically evaluate existing methodologies, identify unresolved issues, stimulate research in these areas, and develop guidance to enhance methodological accuracy of real- life data analyses. At present, STRATOS involves >100 researchers with expertise in statistical and epidemiological methods, from 19 countries worldwide, who work in 9 topic groups (TGs) and 11 panels. Each TG includes experts in a specific area of statistical research, who often have *not* worked together before and may represent different views and/or favor different methods but try to reach consensus through internal discussions and new joint projects. Furthermore, by promoting inter-TG collaborations, STRATOS creates opportunities for developing new interdisciplinary approaches to complex analytical challenges that require cutting-edge expertise in various areas of modern statistics.

In this spirit, the proposed session will involve speakers from 6 STRATOS TGs who will discuss a range of methodological challenges related to different stages of predictive modelling. Specifically, the 4 talks will address: (i) the need for careful initial data analysis to enhance the accuracy and reproducibility of prediction; (ii) the choice of the analytical approach for prediction based on high-dimension data; (iii) criteria and methods to validate and evaluate predictions for survival outcomes; (iv) the use of the predicted values as either exposures or outcomes in further statistical analyses.

Comment on the originality of the topic and its relevance to the IBS. Please comment on the make up of your speakers and discussant, if you have one, and how your proposed session will contribute to the overall diversity (regions, gender, and other aspects) of our IBC2022 program.

<u>Relevance to the IBS:</u>
Predictive modelling with observational data is highly relevant for a wide range of core research activities of IBS members, involving both (a) new developments in statistical

methodology and (b) their applications in collaborative projects in health, biology, agriculture, etc. To ensure the relevance of our session for both (a) and (b), each of the 4 talks will present: (a) cutting-edge new analytical results, and (b) a case study that will illustrate the corresponding challenges and applications of the proposed methods. Jointly, the speakers will offer new interdisciplinary perspectives on methodological issues at the cross-roads of predictive modelling with other areas of statistical research: (1) initial data analysis, (2) high-dimensional data and machine learning, (3) survival analysis, and (4) measurement errors.

Originality:
Each talk will include original information, not published and not presented at previous IBC conferences. Specifically, each talk will present original results of methodological research as well as a novel application in health research, and will also offer new practical insights for analysts. To demonstrate the above points, below we present brief summaries of each of the 4 talks.

(1) Initial data analysis to support prediction modelling in observational studies
Presenters: Marianne Huebner (TG3: Initial data analysis) & Georg Heinze (TG2: Variable selection and functional forms)

Statistical model building for prediction in the life sciences inevitably must take into account properties of the collected data that cannot easily be anticipated. Indeed, as the talk will demonstrate, overlooked unexpected data properties, unsystematic exploratory data analysis and/or lack of transparent reporting may threaten the validity and reproducibility of prediction models. We developed a general strategy to screen data before fitting a prediction model. Our approach relies on criteria for data screening that can be integrated in electronic laboratory notebooks (ELN) to improve transparency and reproducibility of methodological decisions made during prediction modelling. Such an initial data analysis supports the modeller by suggesting modifications to the original statistical analysis plan, and by guiding interpretation and presentation without compromising modelling results. We demonstrate the utility of our proposal in two applications involving: diagnostic prediction of bacteraemia with 50 laboratory variables and an international study predicting grip strength in aging populations.

(2) Statistical and machine learning techniques: relative advantages and weaknesses for prediction with high-dimensional data
Presenters: Jörg Rahnenführer (TG9: High-dimensional data) & Lara Lusa (TG9)

For predictive modelling related to human diseases, machine learning techniques are becoming more and more popular, supported by some spectacular results of deep learning techniques in applications with large numbers of observations. However, in medical research, often sample sizes are moderate or even small, especially in comparison to high-dimensional genetic measurements used as covariates or predictors. Furthermore, predictions are typically associated with uncertainty that must be taken into account. Indeed, the

uncertainty is typically an intrinsic part of the prediction because the further evolution of a disease is not known at the time the prediction is made.

From this perspective, we compare the relative advantages and disadvantages of machine learning approaches *and* more traditional (parametric or quasi-parametric) statistical modelling. We also discuss when and how much to trust the claimed success of a prediction method, especially from a medical perspective, and provide recommendations for evaluating prediction methods. These issues will be illustrated using the example of cancer classification with high-dimensional genetic measurements.

(3) Assessing performance of survival predictions models
Presenters: David McLernon (TG6: Predictive models) & Terry Therneau (TG8: Survival analysis)

Risk prediction models need reliable validation metrics to understand their performance. Methods for the validation of models for survival outcomes have been proposed but many do not appropriately address challenges due to censoring of observations and the varying time horizon at which predictions can be made. We aim to give a hands-on description of methods to evaluate predictions and decisions from survival models based on Cox proportional hazards regression, the most widely used model in this area. We will discuss statistical measures of performance in terms of discrimination, calibration and overall performance, and provide guidance on methodologically sound measures for evaluating predictions and decisions from survival models. As a case study we will discuss prediction of event free survival in breast cancer patients.

(4) Cautionary notes for regression analyses that use a predicted value as either an outcome or an exposure
Presenters: Pamela Shaw (TG4: Measurement Errors) & Laurence Freedman (TG4)

When an important variable in an observational study is hard to measure, an appealing strategy is to *predict* its values from other variables. In essence, this induces Berkson measurement error, the implications of which may not be widely understood. We discuss two scenarios: when the variable for which the predicted values are used is, respectively, an exposure variable and an outcome variable in a regression model, with the exposure-outcome association being of interest. We show that both the impacts of the Berkson measurement error, and appropriate mitigation strategies, differ between these scenarios, and that the assumption of non-differential measurement error plays a central role. We illustrate these issues with data from the Hispanic Community Health Study / Study of Latinos (HCHS/SOL).

Diversity:

We have decided to nominate 2 Speakers for each of the 4 talks, who will prepare and give their presentation *jointly*. All 8 Speakers confirmed their willingness to attend. With this

slightly unorthodox (but not uncommon for STRATOS sessions) approach, we can ensure that each topic is elucidated from 2 different methodological perspectives. Moreover, our approach further enhances the diversity and balance across geographic regions and gender. Specifically, the 8 Speakers include 3 women and 5 men, from 6 countries on 3 continents, and – by representing 6 different STRATOS topic groups (TGs) – offer a truly interdisciplinary perspective on the methodological challenges related to prediction. All Speakers are internationally recognized experts in their fields of statistical research.

Topics Covered and Proposed Speakers/Discussant:

	Affiliation	Contact address	Email address	Tentative title for talk
Marianne Huebner	Department of Statistics and Probability, Michigan State University, East Lansing, MI, USA	619 Red Cedar Rd, East Lansing, MI 48824, USA	huebner@msu.edu	Initial data analysis to support prediction modelling in observational studies
Georg Heinze	Section for Clinical Biometrics, Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Vienna, Austria	Spitalgasse 23, 1090 Vienna, Austria	georg.heinze@meduniwien.ac.at	Initial data analysis to support prediction modelling in observational studies
Jörg Rahnenführer	Department of Statistics, TU Dortmund University, Dortmund, Germany	Vogelpothsweg 87, 44221 Dortmund, Germany	rahnenfuehrer@statistik.tu-dortmund.de	Statistical and machine learning techniques: relative advantages and weaknesses for prediction with high-dimensional data
Lara Lusa	Department of Mathematics, Faculty for Mathematics, Natural Sciences and Information Technologies, University of Primorska, Koper/Capodistria, Slovenia Institute for Biostatistics and Medical Informatics, Medical Faculty,	Galgoljaska 8, 6000 Koper/Capodistria, Slovenia	lara.lusa@famnit.upr.si	Statistical and machine learning techniques: relative advantages and weaknesses for prediction with high-dimensional data

	University of Ljubljana, Slovenia			
David McLernon	Medical Statistics Team, Institute of Applied Health Sciences, University of Aberdeen, Foresterhill, Aberdeen, UK	Robbies Brig, Slains, Ellon, Aberdeenshire, AB41 8SP, UK	d.mclernon@abdn.ac.uk	Assessing performance of survival predictions models
Terry Therneau	Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester MN, USA	Department of Quantitative Health Sciences, Mayo Clinic, 200 First Street SW, Rochester, Minnesota, 55902, USA	therneau@mayo.edu	Assessing performance of survival predictions models
Pamela Shaw	Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Perelman School of Medicine, Philadelphia, USA	606 Blockley Hall, Philadelphia, PA 19104, USA	shawp@upenn.edu	Cautionary notes for regression analyses that use a predicted value as either an outcome or an exposure
Laurence Freedman	Biostatistics and Biomathematics Unit, Gertner Institute for Epidemiology and Health Policy Research, Sheba Medical Center, Tel Hashomer, Israel	Tel Hashomer 52621, Israel	lsf@actcom.co.il	Cautionary notes for regression analyses that use a predicted value as either an outcome or an exposure

Discussant	Affiliation	Contact address	Email address

I confirm that all proposed speakers and discussants have been consulted and have expressed their willingness to participate in this Invited Session, if accepted.

Further comments:

Since 2017 all 9 topic groups (TGs) and 3 panels of the STRATOS initiative published a series of short articles in Volumes 34-37 of the *Biometric Bulletin*. Each group discussed the main issues, concepts, current work and aims for the (near) future.