# 67. BIOMETRISCHES KOLLOQUIUM

## „Tatort Biometrie“ – „Scenes from Biostatistics“

Foto: Presseamt Münster / Angelika Klauser

## 14–17 March 2021

Westfälische Wilhelms-Universität Münster

# ABSTRACTS

Editors

Dominic Enders
Dennis Görlich
Raphael Koch
Rene Schmidt
Carsten Szardenings

IBS

WWU MÜNSTER

IBKF
Institut für Biometrie und Klinische Forschung

Layout by Oleg Choubine

# Contents

**Contributed Talk**

# Statistical humor in classroom: Jokes and cartoons for significant fun with relevant effect

**Annette Aigner**
Charité Universitätsmedizin Berlin, Germany

Small talk with a statistician: Q: “What’s your relationship with your parents? A: “1:2”

Such and similar, short or long jokes, but also cartoons and other humorous means not only amuse statisticians, but also create an easy, positive access for students to a subject generally perceived as difficult, such as statistics.

This article aims to highlight the relevance and positive effects of humor in teaching in general, but especially of easy-to-use materials such as jokes and cartoons. Hints and suggestions for their proper use are given, but of course there are no limits to their implementation in the classroom. In addition, the article contains a collection of freely available online resources that can be used immediately in the statistics classroom and in which everyone can find materials suitable for the specific teaching situation. As a exemplary application, materials for use in an introductory session on linear regression are shown and the author’s personal experiences are briefly summarized.

As statisticians, we know that statistics is fun - now we should also convey this to students, why not with the help of jokes and cartoons?

**Contributed Talk**

# Valid sample size re-estimation at interim

**Nilufar Akbari**
Charité - Institute of Biometry and Clinical Epidemiology, Germany

Throughout this work, we consider the situation of a two-arm controlled clinical trial based on time-to-event data.

The aim of this thesis is to estimate a meaningful survival model in a robust way to observed data during an interim analysis in order to carry out a valid sample size recalculation.

Adaptive designs provide an attractive possibility of changing study design parameters in an ongoing trial. There are still many open questions with respect to adaptive designs for time-to-event data. Among other things, this is because survival data, unlike continuous or binary data, undertake a follow-up phase, so that the outcome is not directly observed after patient's treatment.

Evaluating survival data at interim analyses leads to a patient overrun since the recruitment is usually not stopped at the same time. Another problem is that there must be an interim analysis during this recruitment phase to save patients. Moreover, the timing of the interim analysis is a crucial point build decisions upon a reasonable level of information.

A general issue about time-to-event data is that at an interim analysis one can only calculate the updated size of the required number of events. However, there is normally a greater need in the determination of the sample size to achieve that required number of events. Therefore, the underlying event-time distribution is needed, which may possibly be estimated from the interim data.

This however, is a difficult task for the following reasons: The number of observed events at interim is limited, and the survival curve at interim is truncated by the interim time point.

The goal of this research work is to fit a reasonable survival model to the observed data in a robust way. The fitted curve has the following advantages: The underlying hazards per group can be estimated which allows updating the required number of patients for achieving the respective number of events. Finally, the impact of overrun can be directly assessed and quantified.

The following problems were additionally evaluated in detail. How much do the hazards deviate if the wrong event-time distribution was estimated? At which point in time is a sample size re-estimation useful, or rather how many events are required, for a valid sample size re-estimation at interim?

**Contributed Talk**

# Examining the causal mediating role of brain pathology on the relationship between subclinical cardiovascular disease and cognitive impairment: The Cardiovascular Health Study

**Ryan M. Andrews**[1], **Vanessa Didelez**[1], **Ilya Shpitser**[2], **Michelle C Carlson**[2]
[1]Leibniz Institute for Prevention Research and Epidemiology - BIPS, Germany; [2]Johns Hopkins University

Accumulating evidence suggests that there is a link between subclinical cardiovascular disease and the onset of cognitive impairment in later life. Less is known about possible causal mechanisms underlying this relationship; however, a leading hypothesis is that brain biomarkers play an intermediary role. In this study, we aimed to estimate the proportion of the total effect of subclinical cardiovascular disease on incident cognitive impairment that is mediated through two brain biomarkers--brain hypoperfusion and white matter disease. To do this, we used data from the Cardiovascular Health Study, a large longitudinal cohort study of older adults across the United States. Because brain hypoperfusion and white matter disease may themselves be causally linked with an uncertain temporal ordering, we could not use most multiple mediator methods because we did not believe their assumptions would be met (i.e., that we had independent and causally ordered mediators). We overcame this challenge by applying an innovative causal mediation method—inverse odds ratio weighting—that can accommodate multiple mediators regardless of their temporal ordering or possible effects on each other.

We found that after imposing inclusion and exclusion criteria, approximately 20% of the effect of subclinical cardiovascular disease on incident cognitive impairment was jointly mediated by brain hypoperfusion and white matter disease. We also found that the mediated proportion varied by the type of cognitive impairment, with 21% of the effect being mediated among those with Mild Cognitive Impairment and 12% being mediated among those with dementia.

Interpreting our results as causal effects relies on the plausibility of many assumptions and must be done carefully. Based on subject matter knowledge and the results of several sensitivity analyses, we conclude that most (if not all) assumptions are indeed plausible; consequently, we believe our findings support the idea that brain hypoperfusion and white matter disease are on the causal pathway between subclinical cardiovascular disease and cognitive impairment, particularly Mild Cognitive Impairment. To our knowledge, our study is the first epidemiological study to support the existence of this etiological mechanism. We encourage future studies to extend and to replicate these results.

**Contributed Talk**

# Standardisierte Mittelwertdifferenzen aus Mixed Model Repeated Measures – Analysen

**Lars Beckmann, Ulrich Grouven, Guido Skipka**
Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG), Deutschland

In klinischen Studien werden für Patientinnen und Patienten häufig Daten zur gesundheitsbezogenen Lebensqualität und zur Symptomatik zu aufeinanderfolgenden Zeitpunkten erhoben. Für die Auswertung dieser longitudinalen Daten werden in der Literatur lineare gemischte Modelle für Messwiederholungen (Mixed Models Repeated Measures – Modelle [MMRM]) vorgeschlagen. Diese Endpunkte werden in der Regel mit Skalen mit nicht natürlichen Einheiten gemessen.

Es liegt nahe, für die Bestimmung einer klinischen Relevanz oder für die Durchführung von Metaanalysen auf standardisierte Mittelwertdifferenzen (SMD), wie beispielsweise Cohens d oder Hedges' g, zurückzugreifen. Allerdings ist unklar, wie die für die SMD benötigte gepoolte Standardabweichung aus MMRM – Analysen berechnet werden kann. Anhand einer Simulationsstudie wurden verschiedene Verfahren zur Schätzung einer SMD untersucht. Die Verfahren lassen sich unterteilen in Ansätze, die auf die im MMRM geschätzten Standardfehler der Mittelwertdifferenz (MD) zurückgreifen, und in Ansätze, die die individuellen Patientendaten (IPD) benutzen.

Simuliert wurden Daten einer randomisierten kontrollierten Studie. Die longitudinalen Daten wurden mittels eines autoregressiven Modells 1. Ordnung (AR) für die Abhängigkeiten zwischen den Erhebungszeitpunkten simuliert. Parameter für die Simulationen waren die SMD, die Varianz für die Änderung zum Ausgangswert, die Korrelation für das AR sowie die Stichprobengrößen in den Therapiearmen. Der betrachtete Endpunkt ist die Differenz zwischen den Therapiearmen hinsichtlich der mittleren Änderung zum Ausgangswert über den gesamten Studienverlauf. Die verschiedenen Verfahren wurden bezüglich Überdeckungswahrscheinlichkeit, Verzerrung, Mean Squared Error (MSE), Power und Fehler 1. Art sowie Konkordanz von MD und SMD bez. der statistischen Signifikanz und der Überdeckung des wahren Effektes verglichen.

Die Verfahren, bei denen die gepoolte Standardabweichung aus Standardfehlern des MMRM berechnet wird, zeigen Verzerrungen, die zu einer deutlichen Überschätzung des wahren Effektes führen. Verfahren, die die gepoolte Standardabweichung aus den beobachteten Veränderungen zum Studienanfang schätzen, zeigen eine deutlich geringere Verzerrung und einen geringeren MSE. Allerdings ist die Power, im Vergleich zur MD, kleiner.

Die Schätzung einer SMD mittels der Standardfehler aus dem MMRM ist nicht angemessen. Dies ist insbesondere bei der Bewertung von großen SMDs zu berücksichtigen. Zu einer angemessenen Schätzung einer SMD sind Verfahren notwendig, aus denen die gepoolte Standardabweichung der Änderung zum Ausgangswert mit IPD geschätzt werden kann.

**Contributed Talk**

# Assessment of methods to deal with delayed treatment effects in immunooncology trials with time-to-event endpoints

**Rouven Behnisch, Johannes Krisam, Meinhard Kieser**
Institute of Medical Biometry and Informatics, University of Heidelberg, Germany

In cancer drug research and development, immunotherapy plays an ever more important role. A common feature of immunotherapies is a delayed treatment effect which is quite challenging when dealing with time-to-event endpoints [1]. In case of time-to-event endpoints, regulatory authorities often require a log-rank test, which is the standard statistical method. The log-rank test is known to be most powerful under proportional-hazards alternatives but suffers a substantial loss in power if this assumption is violated. Hence, a rather long follow-up period is required to detect a significant effect in immunooncology trials. For that reason, the question arises whether methods exist that are more susceptible to delayed treatment effects and that can be applied early on to generate evidence anticipating the final decision of the log-rank test to reduce the trial duration without inflation of the type I error. Alternative methods include, for example, weighted log-rank statistics with weights that can either be fixed at the design stage of the trial [2] or chosen based on the observed data [3] or tests based on the restricted mean survival time [4], survival proportions, accelerated failure time (AFT) models or additive hazard models.

We evaluate and compare these different methods systematically with regard to type I error control and power in the presence of delayed treatment effects. Our simulation study includes aspects such as different censoring rates and types, different times of delay, and different failure time distributions. First results show that most methods achieve type I error rate control and that, by construction, the weighted log-rank tests which place more weight on late time points have a greater power to detect differences when the treatment effect is delayed. It is furthermore investigated whether and to what extent these methods can be applied at an early stage of the trial to predict the decision of the log-rank test later on.

**References:**

[1] T. Chen (2013): Statistical issues and challenges in immuno-oncology. Journal for ImmunoTherapy of Cancer 1:18

[2] T.R. Fleming and D.P. Harrington (1991): Counting Processes and Survival Analysis. New York [u.a.]: Wiley-Interscience Publ.

[3] D. Magirr and C. Burman (2019): Modestly weighted logrank tests. Statistics in Medicine 38(20):3782-3790.

[4] P. Royston and M.K.B. Parmar (2013): Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. BMC Medical Research Methodology 13(1):152.

**Contributed Talk**

# Multivariate regression modelling with global and cohort-specific effects in a federated setting with data protection constraints

**Max Behrens, Daniela Zöller**
University of Freiburg, Germany

Multi-cohort studies are an important tool to study effects on a large sample size and to identify cohort-specific effects. Thus, researchers would like to share information between cohorts and research institutes. However, data protection constraints forbid the exchange of individual-level data between different research institutes. To circumvent this problem, only non-disclosive aggregated data is exchanged, which is often done manually and requires explicit permission before transfer. The framework DataSHIELD enables automatic exchange in iterative calls, but methods for performing more complex tasks such as federated optimisation and boosting techniques are missing.

We propose an iterative optimization of multivariate regression models which condenses global (cohort-unspecific) and local (cohort-specific) predictors. This approach will be solely based on non-disclosive aggregated data from different institutions. The approach should be applicable in a setting with high-dimensional data with complex correlation structures. Nonetheless, the amount of transferred data should be limited to enable manual confirmation of data protection compliance.

Our approach implements an iterative optimization between local and global model estimates. Herein, the linear predictor of the global model will act as a covariate in the local model estimation. Subsequently, the linear predictor of the updated local model is included in the global model estimation. The procedure is repeated until no further model improvement is observed for the local model estimates. In case of an unknown variable structure, our approach can be extended with an iterative boosting procedure performing variable selection for both the global and local model.

In a simulation study, we aim to show that our approach improves both global and local model estimates while preserving the globally found effect structure. Furthermore, we want to demonstrate the approach to grant protected access to a multi-cohort data pool concerning gender sensitive studies. Specifically, we aim to apply the approach to improve upon cohort-specific model estimates by incorporating a global model based on multiple cohorts. We will apply the method to real data obtained in the GESA project, where we combined data from the three large German population-based cohorts GHS, SHIP, and KORA to identify potential predictors for mental health protectories.

In general, all gradient-based methods can be adapted easily to a federated setting under data protection constraints. The here presented method can be used in this setting to perform iterative optimisation and can thus aid in the process of understanding cohort-specific estimates. We provide an implementation in the DataSHIELD framework.

**Gustav-Adolf-Lienert Laureate**

# Discrete Subdistribution Hazard Models

**Moritz Berger**
Department of Medical Biometry, Informatics and Epidemiology, Rheinische Friedrich-Wilhelms-Universtät Bonn, Germany

In many clinical and epidemiological studies the interest is in the analysis of the time $T$ until the occurrence of an event of interest $j$ that may occur along with one or more competing events. This requires suitable techniques for competing risks regression. The key quantity to describe competing risks data is the cumulative incidence function, which is defined in terms of the probability of experiencing $j$ at or before time $t$.

A popular modeling approach for the cumulative incidence function is the proportional subdistribution hazard model by Fine and Gray (1999), which is a direct modeling approach for the cumulative incidence function of one specific event of interest. A limitation of the subdistribution hazard model is that it assumes continuously measured event times. In practice, however, the exact (continuous) event times are often not recorded. Instead, it may only be known that the events occurred between pairs of consecutive points in time (i.e., within pre-specified follow-up intervals). In these cases, time is measured on a discrete scale.

To address this issue, a technique for modeling subdistribution hazards with right-censored data in discrete time is proposed. The method is based on a weighted maximum likelihood estimation scheme for binary regression and results in consistent and asymptotically normal estimators of the model parameters. In addition, a set of tools to assess the calibration of discrete subdistribution hazard models is developed. They consist of a calibration plot for graphical assessments as well as a recalibration model including tests on calibration-in-the-large and refinement.

The modeling approach is illustrated by an analysis of nosocomial pneumonia in intensive care patients measured on a daily basis.

**Contributed Talk**

# Der Lernzielkatalog Medizinische Biometrie für das Studium der Humanmedizin

**Ursula Berger**[1], **Carolin Herrmann**[2]
[1]LMU München, Germany; [2]Charité - Universitätsmedizin Berlin, Germany

Der Lernzielkatalog Medizinische Biometrie für das Studium der Humanmedizin umfasst zentrale biometrische Begriffe, Kennzahlen, Konzepte und Methoden sowie Fertigkeiten, die Medizinstudierenden ein Grundverständnis für Biometrie und Datenanalyse vermitteln. Er soll die Planung von Lehrangeboten zur Medizinischen Biometrie im Studium der Humanmedizin erleichtern und Studierenden eine Orientierungshilfe bieten.

Der Lernzielkatalog listet die verschiedenen Lernthemen nach Oberthemen zusammengefasst auf. Zu jedem Lernthema werden die geforderten Fähigkeiten, Fertigkeiten und Kenntnisse der Studierenden durch Verben beschrieben, die auch den Wissensgrad bzw. die Ebene der Lernziele widerspiegelt. Zusätzlich wurden die Lernthemen mit Anmerkungen und Hinweisen für die Lehrenden ergänzt. Bei der Erstellung der Lernthemen wurde der neue Nationale Kompetenzbasierte Lernzielkatalog Medizin NKLM 2.0 im aktuell verfügbaren Entwicklungsstadium (11.2020) berücksichtigt. Der Lernzielkatalog gibt keine Abfolge und keinen zeitlichen Rahmen für ein Curriculum vor und kann daher flexibel in unterschiedlich strukturierten Curricula und unterschiedlichen Typen von Studiengängen der Humanmedizin angewendet werden.

Die Erstellung des Lernzielkatalogs wurde von der gemeinsamen Arbeitsgruppe Lehre und Didaktik der Biometrie der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (GMDS) und der Internationalen Biometrischen Gesellschaft der Deutschen Region (IBS-DR) koordiniert. Dazu wurden in 2020 mehrere Workshops ausgerichtet, in welchen unter der Mitwirkung vieler Fachkolleg*innen eine erste Version erarbeitet werden konnte, die im Dezember 2020 der Fachöffentlichkeit zur Kommentierung vorgestellt wurde (send-to: LZK-Biometrie@charite.de).

Der Lernzielkatalog Medizinische Biometrie für das Studium der Humanmedizin soll nun, nach Ender der Kommentierungsphase, in seiner überarbeiteten Version vorgestellt werden.

**Contributed Talk**

# The key distinction between Association and Causality exemplified by individual ancestry proportions and gallbladder cancer risk in Chileans

**Justo Lorenzo Bermejo, Linda Zollner**
Statistical Genetics Research Group, Institute of Medical Biometry and Informatics, University of Heidelberg, Germany

**Background:**

The translation of findings from observational studies into improved health policies requires further investigation of the type of relationship between the exposure of interest and particular disease outcomes. Observed associations can be due not only to underlying causal effects, but also to selection bias, reverse causation and confounding.

As an example, we consider the association between the proportion of Native American ancestry and the risk of gallbladder cancer (GBC) in genetically admixed Chileans. Worldwide, Chile shows the highest incidence of GBC, and the risk of this disease has been associated with the individual proportion of Native American – Mapuche ancestry. However, Chileans with large proportions of Mapuche ancestry live in the south of the country, have poorer access to the health system and could be exposed to distinct risk factors. We conducted a Mendelian Randomization (MR) study to investigate the causal relationship “Mapuche ancestry → GBC risk”.

**Methods:**

To infer the potential causal effect of specific risk factors on health-related outcomes, MR takes advantage of the random inheritance of genetic variants and utilizes instrumental variables (IVs):

1. associated with the exposure of interest
2. independent of possible confounders of the association between the exposure and the outcome
3. independent of the outcome given the exposure and the confounders

Given the selected IVs meet the above assumptions, various MR approaches can be used to test causality, for example the inverse variance weighted (IVW) method.

In our example, we took advantage of ancestry informative markers (AIMs) with distinct allele frequencies in Mapuche and other components of the Chilean genome, namely European, African and Aymara-Quechua ancestry. After checking that the AIMs fulfilled the required assumptions, we utilized them as IVs for the individual proportion of Mapuche ancestry in two-sample MR (sample 1: 1,800 Chileans from the whole country, sample 2: 250 Chilean case-control pairs).

**Results:**

We found strong evidence for a causal effect of Mapuche ancestry on GBC risk: IVW OR per 1% increase in the Mapuche proportion 1.02, 95%CI (1.01-1.03), Pval = 0.0001. To validate this finding, we performed several sensitivity analyses including radial MR and different combinations of genetic principal components to rule out population stratification unrelated to Mapuche ancestry.

**Conclusion:**

Causal inference is key to unravel disease aetiology. In the present example, we demonstrate that Mapuche ancestry is causally linked to GBC risk. This result can now be used to refine GBC prevention programs in Chile.

**Invited Talk**

# Do we still need hazard ratios? (II)

**Jan Beyersmann**
Ulm University, Germany

The answer to the question whether we need hazard ratios depends to a good deal on the answer to the question what we need hazards for. Censoring plays a key role. Censoring makes survival and event history analysis special. One important consequence is that less customized statistical techniques will be biased when applied to censored data. Another important consequence is that hazards remain identifiable under rather general censoring mechanisms.

In this talk, I will demonstrate that there is a Babylonian confusion on “independent censoring” in the textbook literature, which is a worry in its own right. Event-driven trials in pharmaceutical research or competing risks are two examples where the textbook literature often goes haywire, censoring-wise. It is a small step from this mess to misinterpretations of hazards, a challenge not diminished when the aim is a causal interpretation. Causal reasoning, however, appears to be spearheading the current attack on hazards and their ratios.

In philosophy, causality has pretty much been destroyed by David Hume. This does not imply that statisticians should avoid causal reasoning, but it might suggest some modesty. In fact, statistical causality is mostly about interventions, and a causal survival analysis often aims at statements about the intervention “do(no censoring)”, which, however, is not what identifiability of hazards is about. The current debate about estimands (in time-to-event trials) is an example where these issues are hopelessly mixed up.

The aim of this talk is to mix it up a bit further or, perhaps, even shed some light. Time permitting, I will illustrate matters using g-computation in the form of a causal variant of the Aalen-Johansen-estimator.

**Contributed Talk**

# Evaluation of event rate differences using stratified Kaplan-Meier difference estimates with Mantel-Haenszel weights

**Hannes Buchner**[1], **Stephan Bischofberger**[1], **Rainer-Georg Goeldner**[2]
[1]Staburo, Germany; [2]Boehringer Ingelheim, Germany

The assessment of differences in event rates is a common endeavor in the evaluation of the efficacy of new treatments in clinical trials. We investigate the performance of different hypothesis tests for cumulative hospitalization or death rates of Covid-19 in order to reliably determine the efficacy of a novel treatment. The focus of the evaluation was on the comparison of event rates via Kaplan-Meier estimates for a pre-specified day and the aim to reduce sampling error, hence we examine different stratum weights for a stratified Z-test for Kaplan-Meier differences. The simulated data is calibrated from recent research on neutralizing antibodies treatment for Covid-19 with 2, 4, and 6 strata of different size and prevalence, and we investigate the effects of overall event rates ranging from 2% to 20%. We simulate 1000 patients and compare the results of 1000 simulation runs. Our simulation study shows superior performance of Mantel-Haenszel-type weights [Greenland & Robins (1985), Biometrics 41, 55-68] over inverse variance weights – in particular for unequal stratum sizes and very low event rates in some strata as common in COVID-19 treatment studies. The advantage of this approach is a larger power of the test (e.g. 79% instead of 64% for an average event rate 7%). The results are compared with those of a Cochran-Mantel-Haenszel (CMH) test, which yields lower power than the inverse variance weights for low event rates (under 62% for average event rate 7%) and consistently lower power than the Z-test with the Mantel-Haenszel stratum weights. Moreover, the CMH test breaks down (power reduction by 30%) in presence of loss-to-follow-up with as little as 5% of the patients due to its nature of not being designed for time-to-event data. The performance of the Z-test for Kaplan-Meier differences on the other hand is hardly suffering from the latter (power reduction by 4%). All investigated tests satisfy the set significance level for type-I errors in our simulation study.

**Contributed Talk**

# Tumour-growth models improve progression-free survival estimation in the presence of high censoring-rates

**Gabriele Bleckert, Hannes Buchner**
Staburo GmbH, Germany

**Introduction:**

In oncology, reliable estimates of progression-free survival (PFS) are of highest importance because of high failure rates of phase III trials (around 60%). However, PFS estimations on early readouts with less than 50% of events observed do not use all available information from tumour measurements over time.

**Method:**

We project the PFS-event of each censored patient by using a mixed model [2] describing the tumour burden over time. RECIST-criteria are applied on estimated patient-specific non-linear tumour-trajectories to calculate the projected time-to-progression.

PFS is compared between test and reference by hazard ratios (HR).

Several phase III and II simulations with 1000 runs each with 2000 or 80 patients, 6 months accrual and 2 (scenario-1) or 6 months (scenario-2) follow-up were performed.

All simulations are based on a published optimal parameterisation [1] of tumour-growth in non-small-cell lung cancer (NSCLC) which implies a time-dependent HR.

**Results:**

The classical PFS estimation resulted in a HR of 0.34 (95%-percentiles: 0.29-0.40) for scenario-1 and 0.52 (0.47-0.58) for scenario-2 compared to a predicted HR of 0.77 for both scenarios (0.69-0.85 and 0.69-0.84), while the overall true HR (over ten years) was 0.78 (0.69-0.85). For 6, 12 and 120 months the time varying HRs were 0.41 (0.36-0.47), 0.60 (0.54-0.66) and 0.77 (0.69-0.85).

The classical PFS estimation for phase II showed HRs from 0.52 to 0.61 compared to predicted HRs between 0.71 and 0.77.

**Conclusion:**

Tumour-growth models improve PFS estimations in the presence of high censoring-rates as they consistently provide far better estimates of the overall true HR in phase III and II trials.

**References:**

[1] M. Reck, A. Mellemgaard, S. Novello, PE. Postmus, B. Gaschler-Markefski, R. Kaiser, H. Buchner: Change in non-small-cell lung cancer tumor size in patients treated with nintedanib plus docetaxel: analyses from the Phase III LUME-Lung 1 study, OncoTargets and Therapy 2018:11 4573–4582

[2] Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics. 1982 Dec;38(4):963-74. PMID: 7168798.

**Contributed Talk**

# Sample size calculation and blinded re-estimation for diagnostic accuracy studies considering missing or inconclusive test results

**Cordula Blohm**[1], **Peter Schlattmann**[2], **Antonia Zapf**[3]
[1]Universität Heidelberg; [2]Universitätsklinikum Jena; [3]Universitätsklinikum Hamburg-Eppendorf

**Background:**

For diagnostic accuracy studies the two independent co-primary endpoints, sensitivity and specificity, are relevant. Both parameters are calculated based on the disease status and the test result that is evaluated either as positive or negative. Sometimes the test result is neither positive nor negative but inconclusive or even missing. There are four frequently used methods of handling missing values available where such results are counted as missing, positive, negative, or false positive and false negative. The first three approaches may lead to an overestimation of both parameters, or either sensitivity or specificity, respectively. In the fourth approach, the intention to diagnose principle (ITD), both parameters decrease and a more realistic picture of the clinical potential of diagnostic tests is provided (Schuetz et al. 2012).

Sensitivity and specificity are also key parameters in sample size calculation of diagnostic accuracy studies and a realistic estimate of them is mandatory for the success of a trial. Therefore, the consideration of inconclusive results in the initial sample size calculation and, especially, in a blinded sample size re-calculation based on an interim analysis could improve trial design.

**Methods:**

For sample size calculation, the minimum sensitivity and specificity, the type I error rate and the power are defined. In addition, the expected sensitivity and specificity of the experimental test, the prevalence, and the proportion of inconclusive results are assumed. For the simulation study different scenarios are chosen by varying these parameters. The optimal sample size is calculated according to Stark and Zapf (2020). The inconclusive results are generated independently of disease status and randomly distributed over diseased and non-diseased subjects. The sensitivity and specificity of the experimental test are estimated while considering the four different methods, mentioned above, to handle inconclusive results.

The sample size re-recalculation is performed with a blinded one-time re-estimation of the proportion of inconclusive results. The power, the type I error rate, and the bias of estimated sensitivity and specificity are used as performance measures.

**Results:**

The simulation study aims to evaluate the influence of inconclusive results on the evaluation of diagnostic test accuracy in an adaptive study design. The performance difference of the four methods to handle inconclusive results will be discussed.

**Contributed Talk**

# A replication crisis in methodological statistical research?

**Anne-Laure Boulesteix[1], Stefan Buchka[1], Alethea Charlton[1], Sabine Hoffmann[1], Heidi Seibold[2], Rory Wilson[2]**
[1]LMU Munich, Germany; [2]Helmholtz Zentrum Munich, Germany

Statisticians are often keen to analyze the statistical aspects of the so-called “replication crisis”. They condemn fishing expeditions and publication bias across empirical scientific fields applying statistical methods. But what about good practice issues in their own - methodological - research, i.e. research considering statistical methods as research objects? When developing and evaluating new statistical methods and data analysis tools, do statisticians adhere to the good practice principles they promote in fields which apply statistics? I argue that statisticians should make substantial efforts to address what may be called the replication crisis in the context of methodological research in statistics and data science. In the first part of my talk, I will discuss topics such as publication bias, the design and necessity of neutral comparison studies and the importance of appropriate reporting and research synthesis in the context of methodological research.

In the second part of my talk I will empirically illustrate a specific problem which affects research articles presenting new data analysis methods. Most of these articles claim that “the new method performs better than existing methods”, but the veracity of such statements is questionable. An optimistic bias may arise during the evaluation of novel data analysis methods resulting from, for example, selection of datasets or competing methods; better ability to fix bugs in a preferred method; and selective reporting of method variants. This bias is quantitatively investigated using a topical example from epigenetic analysis: normalization methods for data generated by the Illumina HumanMethylation450K BeadChip microarray.

**Contributed Talk**

# Personalstruktur und Outcome in der stationären Langzeitpflege – Methoden und Limitationen einer statistischen Auswertung von longitudinalen Routinedaten

**Werner Brannath, Pascal Rink**
Institut für Statistik und KKSB, Fachbereich Mathematik und Informatik, Universität Bremen

Angesichts des demographischen Wandels und dem gleichzeitigen Mangel an professionellen Pflegekräften, gewinnt die Frage nach dem benötigten Umfang und der adäquaten Struktur des Pflegepersonals einer Einrichtung mit stationärer Langzeitpflege zunehmend an gesellschaftlicher und politischer Bedeutung. Neben einer vom Gesetzgeber in Auftrag gegebenen Studie zur Entwicklung eines einheitlichen Verfahrens zur Bemessung des Personalbedarfs, wurde dieser Frage in einer auf longitudinale Routinedaten basierenden Beobachtungsstudie (StaVaCare 2.0) nachgegangen. Ziel der letzteren war es, Erkenntnisse über den komplexen Zusammenhang zwischen der Bewohnerstruktur (Care-Mix) und der Qualifikationsstruktur sowie des Personaleinsatzes des Pflegepersonals (Case-Mix) einer Einrichtung in Hinblick auf dessen Pflege-Outcome zu gewinnen. Die Beschränkung auf Routinedaten führte naturgemäß zu Limitationen und Komplikationen bzgl. der Datenqualität und Datendichte, den statistischen Auswertungen und ihrer Interpretation. Sie lieferte aber anderseits die Möglichkeit zur Vollerhebung innerhalb der Einrichtungen. Darüber hinaus gibt es bisher kein allgemein anerkanntes Verfahren zur Erhebung und Beurteilung des Pflege-Outcomes. In diesem Vortrag sollen die in StaVaCare 2.0 gewählten statistischen Ansätze zur Lösung bzw. Abschwächung der genannten Schwierigkeiten beschrieben und diskutiert werden.

**References:**

[1] Görres S, Brannath W, Böttcher S, Schulte K, Arndt G, Bendig J, Rink P, Günay S (2020). Stabilität und Variation des Care-Mix in Pflegeheimen unter Berücksichtigung von Case-Mix, Outcome und Organisationscharakteristika (StaVaCare 2.0). Abschlussbericht des Modellvorhabens mit Anhang.

[2] https://www.gkv-spitzenverband.de/pflegeversicherung/forschung/modellprojekte/pflege_abgeschlossene_projekte_8/stava.jsp

**Contributed Talk**

# Single-stage, three-arm, adaptive test strategies for non-inferiority trials with an unstable reference

**Werner Brannath, Martin Scharpenberg, Sylvia Schmidt**
University of Bremen, Germany

For indications where only unstable reference treatments are available and use of placebo is ethically justified, three-arm `gold standard' designs with an experimental, reference and placebo arm are recommended for non-inferiority trials. In such designs, the demonstration of efficacy of the reference or experimental treatment is a requirement. They have the disadvantage that only little can be concluded from the trial if the reference fails to be efficacious. To overcome this, we investigate a novel single-stage, adaptive test strategies where non-inferiority is tested only if the reference shows sufficient efficacy and otherwise delta-superiority of the experimental treatment over placebo is tested. With a properly chosen superiority margin, delta-superiority indirectly shows non-inferiority. We optimize the sample size for several decision rules and find that the natural, data driven test strategy, which tests with non-inferiority if the reference's efficacy test is significant, leads to the smallest overall and placebo sample sizes. Under specific constraints on the sample sizes, this procedure controls the family-wise error rate. All optimal sample sizes are found to meet this constraint. We finally show how to account for a relevant placebo drop-out rate in an efficient way and apply the new test strategy to a real life data set.

**Contributed Talk**

# Graphical approaches for the control of generalized error rates

**Frank Bretz**[1], **David Robertson**[2], **James James Wason**[3]
[1]Novartis, Switzerland; [2]University of Cambridge, UK; [3]Newcastle University, UK

When simultaneously testing multiple hypotheses, the usual approach in the context of confirmatory clinical trials is to control the familywise error rate (FWER), which bounds the probability of making at least one false rejection. In many trial settings, these hypotheses will additionally have a hierarchical structure that reflects the relative importance and links between different clinical objectives. The graphical approach of Bretz et al (2009) is a flexible and easily communicable way of controlling the FWER while respecting complex trial objectives and multiple structured hypotheses. However, the FWER can be a very stringent criterion that leads to procedures with low power, and may not be appropriate in exploratory trial settings. This motivates controlling generalized error rates, particularly when the number of hypotheses tested is no longer small. We consider the generalized familywise error rate (k-FWER), which is the probability of making k or more false rejections, as well as the tail probability of the false discovery proportion (FDP), which is the probability that the proportion of false rejections is greater than some threshold. We also consider asymptotic control of the false discovery rate, which is the expectation of the FDP. In this presentation, we show how to control these generalized error rates when using the graphical approach and its extensions. We demonstrate the utility of the resulting graphical procedures on clinical trial case studies.

**Contributed Talk**

# Diagnostic accuracy of claims data from 70 million people in the German statutory health insurance: Type 2 diabetes in men

**Ralph Brinks**[1,2,3], **Thaddäus Tönnies**[1], **Annika Hoyer**[2]

[1]Deutsches Diabetes-Zentrum, Germany; [2]Department of Statistics, Ludwig-Maximilians-University Munich; [3]Department of Rheumatology, University Hospital Duesseldorf

During estimation of excess mortality in people with type 2 diabetes in Germany based on aggregated claims data from about 70 million people in the statutory health insurance, we experienced and reported problems in the age groups below 60 years of age [1]. We hypothesized that diagnostic accuracy (sensitivity and specificity) might be the reason for those problems [1].

In the first part of this work, we ran a simulation study to assess the impact of the diagnostic accuracy on the estimation of excess mortality. It turns out that the specificity in the younger age groups has the greatest effect on the estimate in terms of bias of the excess mortality while the sensitivity has a much lower impact.

In the second part, we apply these findings to estimate the diagnostic accuracy of type 2 diabetes in men aged 20-90 based on the approach and data from [1]. We obtain that irrespective of the sensitivity, the false positive ratio (FPR) increases linearly from 0.5 to 2 per mil from age 20 to 50. At ages 50 to 70, the FPR is likely to drop to 0.5 per mil, followed by a steep linear increase to 5 per mil at age 90.

Our examination demonstrates the crucial impact of diagnostic accuracy on estimates based on secondary data. While for other epidemiological measures sensitivity might be more important, estimation of excess mortality crucially depends on the specificity of the data. We use this fact to estimate the age-specific FPR of diagnoses of type 2 diabetes in aggregated claims data.

**Reference:**

[1] Brinks R, Tönnies T, Hoyer A (2020) DOI 10.1186/s13104-020-05046-w

**Contributed Talk**

# Quality control in genome-wide association studies revisited: a critical evaluation of the standard methods

**Hanna Brudermann[1], Tanja K. Rausch[1,2], Inke R. König[1]**

[1]Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Germany, Germany; [2]Department of Pediatrics, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck, Germany

Genome-wide association studies (GWAs) investigating the relationship between millions of genetic markers and a clinically relevant phenotype were originally based on the common disease - common variant assumption, thus aiming at identifying a small number of common genetic loci as cause for common diseases. Given the enormous cost reduction in the acquisition of genomic data, it is not surprising that since the first known GWA by Klein et al. (2005), this study type was established as a standard method. However, since even low error frequencies can distort association results, extensive and accurate quality control of the given data is mandatory. In recent years, the focus of GWAs has shifted, and the task is no longer primarily the discovery of common genetic loci. Also, with increasing sample sizes and (mega-)meta-analyses of GWAs, it is hoped that loci with small effects can be identified. Furthermore, it has become popular to aggregate all genomic information, even loci with very small effects and frequencies, into genetic risk prediction scores, thus increasing the requirement for high-quality genetic data.

However, after extensive discussions about standards for quality control in GWAs in the early years, further work on how to control data quality and adapt data cleaning to new GWAs aims has become scarce.

The aim of this study was to perform an extensive literature review to evaluate currently applied quality control criteria and their justification. Building on the findings from the literature search, a workflow was developed to include justified quality control steps, keeping in mind that a strict quality control, which removes all data with a high risk of bias, always carries the risk that the remaining data is too homogeneous to make small effects visible. This workflow is subsequently illustrated using a real data set.

Our results show that in most published GWAs, no scientific reasons for the applied quality steps are given. Cutoffs for the most common quality measures are mostly not explained. Especially the principal component analysis and the test for deviation from Hardy-Weinberg equilibrium are frequently used as quality criteria in many GWAs without analyzing the existing conditions exactly and adjusting the quality control accordingly.

It is pointed out that researchers still have to decide between universal and individual parameters and therefore between optimal comparability to other analyses and optimal conditions within the specific study.

**Invited Talk**

# Methodische Impulse und statistische Analyseverfahren, die zur Theorieentwicklung und -Prüfung in der Pflegewissenschaft beitragen können

**Albert Bruehl**
Philosophisch-Theologische Hochschule Vallendar

Statistische Analyseverfahren können Impulse zur Theorieentwicklung in der Pflegewissenschaft geben. Methoden hierzu sind die Entwicklung und Prüfung von Hypothesen. In Standardwerken zur Einführung in die Statistik in den Sozialwissenschaften wird ausschließlich die Hypothesenprüfung als Haupt-Aufgabe der Statistik definiert. Hypothesenentwicklung wäre eine zusätzliche Aufgabe für die Anwendung von Statistik, die bei Gegenständen, wie sie in der Pflegewissenschaft behandelt werden, besonders wichtig werden kann (Brühl, Fried, 2020).

Bei vielen Fragestellungen innerhalb der Pflegewissenschaft haben wir es nämlich mit Versuchen zu tun, empirische Gegenstände über Konstrukte zu modellieren. Beispiele hierfür wären die Modelle zu den Konstrukten „Pflegebedürftigkeit“ und „Pflegequalität“. Theoretisch grundgelegt und empirisch unterstützt sind diese Modelle nicht.

Werden nun Regressionen mit klassischen H0-Hypothesentests zur Datenanalyse im Bereich von Konstrukten wie der Pflegebedürftigkeit und der Pflegequalität eingesetzt, lernen wir, dass die Konstrukte, die wir zu Pflegebedürftigkeit und Pflegequalität im Einsatz haben, empirisch wenig hilfreich sind. Das gilt für multivariate Regressionen, die Arbeitszeiten mit Hilfe von Pflegegradkriterien schlecht erklären (Rothgang, 2020), das gilt für nicht-parametrische Regressionen, Multivariate Regression Splines und Mehr-Ebenen-Modelle, die Arbeitszeit mit Bewohner- und auch Organisations-Variablen nicht gut erklären (Brühl, Planer 2019) und das gilt auch für logistische Regressionen (Görres et al.,2017) und logistische Mehr-Ebenen-Analysen (Brühl, Planer, 2019), die Qualitätsindikatoren nicht gut erklären. Meist werden trotz der bescheidenen Erfolge der statistischen Analysen, auf dieser Basis trotzdem Anwendungsroutinen z.B. zur Personalbemessung und zur Messung von Pflegequalität etabliert.

Aus dieser Art des Einsatzes von Statistik ergeben sich kaum Ansätze für die Weiterentwicklung der eingesetzten Konstrukte. Hierzu sind strukturierende Verfahren besser geeignet. Beispiel hierfür kann der Einsatz verschiedener Varianten der ordinalen Multidimensionalen Skalierung (Borg, 2018) sein, die bei der Weiterentwicklung des Konstrukts der Pflegebedürftigkeit (Teigeler, 2017) und bei der Erfassung von Prozessqualität (Brühl et al, 2021) helfen. Ein weiteres Verfahren, das hier helfen kann, sind die Multiplen Korrespondenzanalysen (Greenacre, 2017), die auch bei kleinen Fallzahlen und mit Nominaldaten eingesetzt werden können. Zur Theorieprüfung können konfirmatorische Varianten der strukturierenden Verfahren eingesetzt werden. Im Vortrag werden Beispiele hierzu vorgestellt.

**Literatur**

[1] Borg, I., Groenen, P. J., & Mair, P. (2018). Applied multidimensional scaling and unfolding (2nd ed.). Springer-Verlag.

[2] Brühl, A., Planer, K. (2019): PiBaWü - Zur Interaktion von Pflegebedürftigkeit, Pflegequalität und Personalbedarf. Freiburg: Lambertus

[3] Brühl, A. (2020): Anwendung von statistischen Analyseverfahren, die die Entwicklung von Theorien in der Pflegewissenschaft fördern, S. 7 -S. 37. In: Brühl, A., Fried, K. (Hsg.) (2020): Innovative Statistik in der Pflegeforschung. Freiburg: Lambertus

[4] Brühl, A., Sappok-Laue, H., Lau, S., Christ-Kobiela, P., Müller, J., Sesterhenn-Ochtendung, B., Stürmer-Korff, R., Stelzig, A., Lobb, M., Bleidt, W. (2021): Indicating Care Process Quality: A Multidimensional Scaling Analysis. Journal of Nursing Measurement, Volume 30, Number 2, 2021 (Advance online publication) http://dx.doi.org/10.1891/JNM-D-20-00096

[5] Greenacre, M. (2017). Correspondence Analysis in Practice (Third Edition). Chapman & Hall / CRC Interdisciplinary Statistics. Boca Raton: CRC Press Taylor and Francis Group.

[6] Görres, Stefan; Rothgang, Heinz (2017): Modellhafte Pilotierung von Indikatoren in der stationären Pflege (MoPIP). Abschlussbericht zum Forschungsprojekt. (SV14-9015). Unter Mitarbeit von Sophie Horstmann, Maren Riemann, Julia Bidmon, Susanne Stiefler, Sabrina Pohlmann, Mareike Würdemann et al. UBC-Zentrum für Alterns- und Pflegeforschung, UBCZentrumfür Sozialpolitik. Bremen

[7] Rothgang, H., Görres, S., Darmann-Finck, I., Wolf-Ostermann, K., Becke, G, Brannath, W. (2020): Zweiter Zwischenbericht. Online verfügbar unter: https://www.gs-qsa-pflege.de/wp-content/uploads/2020/02/2.-Zwischenbericht-Personalbemessung-%C2%A7-113c-SGB-XI.pdf, zuletzt geprüft am 07.09.2020

[8] Teigeler, Anna Maria. (2017): Die multidimensionale Skalierung als grundlegendes Verfahren zur Explikation des Pflegebedürftigkeitsverständnisses von beruflich Pflegenden. Masterthesis an der Philosophisch Theologischen Hochschule Vallendar.https://kidoks.bsz-bw.de/files/1097/Masterthesis +11+17+V.pdf, letzter Zugriff am 30.05.2019.

**Poster**

# A Scrum related interdisciplinary project between Reproductive Toxicology and Nonclinical Statistics to improve data transfer, statistical strategy and knowledge generation

**Monika Brüning[1], Bernd Baier[2], Eugen Fischer[2], Gaurav Berry[2], Bernd-Wolfgang Igl[1]**

[1]Nonclinical Statistics, Biostatistics and Data Sciences, Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach an der Riss, Germany; [2]Reproductive Toxicology, Nonclinical Drug Safety, Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach an der Riss, Germany

The development of a new drug is a long journey full of important milestones that have to be reached successfully. After identification of a pharmaceutical candidate substance, nonclinical developmental and reproductive toxicology (DART) studies are one important element to assess the safety of the future drug. Recommendations on study design and conduct are given in ICH Guideline S5 to support human clinical trials and market access of new pharmaceuticals involving various phase-dependent designs incl. a huge number of parameters.

DART studies have to be performed in animal models and aim to detect any effect of the test item on mammalian reproduction relevant for human risk assessment.

In general, reproductive toxicology study data involve complex correlation structures between mother and offspring, e.g. maternal weight development, fetus weight, ossification status and number of littermates all in dependence of different test item doses. Thus, from a statistical point of view, DART studies are highly demanding and interesting. This complexity is not reflected in statistical approaches implemented in standard lab software.

To this end, we have developed a Scrum inspired project to intensify the cooperation between Reproductive Toxicology and Nonclinical Statistics to work according to agile principles. Therein, we e.g. defined processes for data transfer and analysis incl. a sophisticated and scientifically state-of-the-art statistical methodology.

In this work, we will mainly focus on technical aspects for constructing an Analysis Data Set (ADS) involving regulatory requirements by CDISC SEND (Standard for Exchange of Nonclinical Data), but also sketch new concepts for visualization and statistical analysis.

**Contributed Talk**

# Discussion of Design and Analysis of Animal Experiments

**Edgar Brunner**
Universitätsmedizin Göttingen, Germany

In this talk, some aspects of particular topics in designing and analysis of animal experiments are discussed. They are important in applications for funding. Below are the keypoints.

- Replication of the Exepriment
  - o Different laboratories
  - o Separate experiments vs. stratification and pooling
  - o Impact of the mother in case of experiments involving young animals
- Randomization and Blinding
  - o Onsite-randomization vs. central randomization
- Sample Size Planning
  - o Quite rarely used
  - o Discussion and definition of a ‘relevant effect’
  - o Effects based observed in a preliminary study – often based on a few observations
  - o Relation to effects in human trials may be a problem
  - o Type-I Error adjusting for multiple / co-primary endpoints
  - o Power adjusting for multiple / co-primary endpoints
  - o Switching from non-normal data to rank procedures
- Analysis
  - o Data base cleaning
  - o Principle ‘analyze as randomized’
  - o Pre-testing assumptions on the same data set is not recommended

**References**

[1] Festing, M.F. (2007). The design of animal experiments. Chpt.3 in: Handbook on Care and Management of Laboratory and Pet Animals. Ed. Y. B. Rajeshwari. ISBN: 8189422987.

[2] Exner, C., Bode, H.-J.,Blumer, K., Giese, C. (2007). Animal Experiments in Research. Deutsche Forschungsgemeinschaft. ISBN 978-3-932306-87-7

**Contributed Talk**

# Assessment of additional benefit for time-to-event endpoints after significant phase III trials – investigation of ESMO and IQWiG approaches

**Christopher Alexander Büsch, Johannes Krisam, Meinhard Kieser**
University of Heidelberg, Germany

New cancer treatments are often promoted as major advances after a significant phase III trial. Therefore, a clear and unbiased knowledge about the magnitude of the clinical benefit of newly approved treatments is important to assess the amount of reimbursement from public health insurance of new treatments. To perform these evaluations, two distinct "additional benefit assessment" methods are currently used in Europe.

The European Society for Medical Oncology (ESMO) developed the Magnitude of Clinical Benefit Scale Version 1.1 (ESMO-MCBSv1.1) classifying new treatments into 5 categories using a dual rule considering the relative and absolute benefit assessed by the lower limit of the 95% HR confidence interval or the observed absolute difference in median treatment outcomes, respectively[1,2]. As an alternative, the German IQWiG compares the upper limit of the 95% HR confidence interval to specific relative risk scaled thresholds classifying new treatments into 6 categories[4]. Until now, these methods have only been compared empirically[3].

We evaluate and compare the two methods in a simulation study with focus on time-to-event outcomes. The simulation includes aspects such as different censoring rates and types, incorrect HRs assumed for sample size calculation, informative censoring, and different failure time distributions. Since no "placebo" method reflecting a true (deserved) maximal score is available, different thresholds of the simulated treatment effects were used as alternatives. The methods' performance is assessed via ROC curves, sensitivity / specificity, and the methods' percentage of achieved maximal scores. Our results indicate that IQWiGs method is usually more conservative than ESMOs. Moreover, in some scenarios such as quick disease progression or incorrect assumed HR IQWiGs method is too liberal compared to ESMO. Nevertheless, further research is required, e.g. methods' performance under non-proportional hazards.

**References:**

[1] N.I. Cherny, U. Dafni et al. (2017): ESMO-Magnitude of Clinical Benefit Scale version 1.1. Annals of Oncology, 28:2340-2366

[2] N.I. Cherny, R. Sullivan et al. (2015): A standardised, generic, validated approach to stratify the magnitude of clinical benefit that can be anticipated from anti-cancer therapies: the European Society for Medical Oncology Magnitude of Clinical Benefit Scale (ESMO-MCBS). Annals of Oncology, 26:1547-1573

[3] U. Dafni, D. Karlis et al. (2017): Detailed statistical assessment of the characteristics of the ESMO Magnitude of Clinical Benefit Scale (ESMO-MCBS) threshold rules. ESMO Open, 2:e000216

[4] G. Skipka, B. Wieseler et al. (2016): Methodological approach to determine minor, considerable, and major treatment effects in the early benefit assessment of new drugs. Biometrical Journal, 58:43-58

**Contributed Talk**

# The augmented binary method for composite endpoints based on forced vital capacity (FVC) in systemic sclerosis-associated interstitial lung disease

**Carolyn Cook[1], Michael Kreuter[2], Susanne Stowasser[3], Christian Stock[4]**

[1]mainanalytics GmbH, Sulzbach, Germany; [2]Center for Interstitial and Rare Lung Diseases, Pneumology and Respiratory Care Medicine, Thoraxklinik, University of Heidelberg, Heidelberg, Germany and German Center for Lung Research, Heidelberg, Germany; [3]Boehringer Ingelheim International GmbH, Ingelheim am Rhein, Germany; [4]Boehringer Ingelheim Pharma GmbH & Co. KG, Ingelheim am Rhein, Germany

### Background

The augmented binary method (Wason & Seaman. Stat Med, 2013; 32(26)) is a novel method for precisely estimating response rates and differences among response rates defined based on a composite endpoint that contains a dichotomized continuous variable and additional inherently binary components. The method is an alternative to traditional approaches such as logistic regression techniques. Due to the complexity and computational demands of the method, experience in clinical studies has been limited thus far and is mainly restricted to oncological studies. Operating characteristics and, thus, potential statistical benefits are unclear for other settings.

### Objective

We aimed to perform a Monte Carlo simulation study to assess operating characteristics of the augmented binary method in settings relevant to randomized controlled trials and non-interventional studies in systemic sclerosis-associated interstitial lung disease (SSc-ILD), a rare, chronic autoimmune disease, where composite endpoints of the above described type are frequently applied.

### Methods

An extensive simulation study was performed assessing type I error, power, coverage, and bias of the augmented binary method and a standard logistic model for the composite endpoint. Parameters were varied to resemble lung function decline (as measured through the forced vital capacity, FVC), hospitalization events and mortality in patients with SSc-ILD over a 1- and 2-year period. A relative treatment effect of 50% on FVC was assumed, while assumed effects on hospitalizations and mortality were derived from joined modeling analyses of existing trial data (as indirect effects of the treatment on FVC). Further, the methods were exemplarily applied to data from the SENSCIS trial, a phase III randomized, double-blind, placebo-controlled trial to investigate the efficacy and safety of nintedanib in patients with SSc-ILD.

### Results

The simulation study is currently in progress and results will be available by the end of January. In preliminary results modest gains in power and precision were observed, with acceptable compromises of type I error, if any. In scenarios with lower statistical powers, these results were more likely to make a difference on inferences concerning the treatment effect. In the exemplary application of the augmented binary method to trial data confidence intervals and p-values on selected endpoints involving FVC decline, hospitalization and mortality were smaller.

### Conclusion

Based on preliminary results from a simulation study, we identified areas where the augmented binary method conveys an appreciable advantage compared to standard methods.

**Contributed Talk**

# Adaptive group sequential designs for phase II trials with multiple time-to-event endpoints

**Moritz Fabian Danzer[1], Tobias Terzer[2], Andreas Faldum[1], Rene Schmidt[1]**

[1]Institute of Biostatistics and Clinical Research, University of Münster, Germany; [2]Division of Biostatistics, German Cancer Research Center, Heidelberg, Germany

Existing methods concerning the assessment of long-term survival outcomes in one-armed trials are commonly restricted to one primary endpoint. Corresponding adaptive designs suffer from limitations regarding the use of information from other endpoints in interim design changes. Here we provide adaptive group sequential one-sample tests for testing hypotheses on the multivariate survival distribution derived from multi-state models, while making provision for data-dependent design modifications based on all involved time-to-event endpoints. We explicitly elaborate application of the methodology to one-sample tests for the joint distribution of (i) progression-free survival (PFS) and overall survival (OS) in the context of an illnessdeath model, and (ii) time to toxicity and time to progression while accounting for death as a competing event. Large sample distributions are derived using a counting process approach. Small sample properties and sample size planning are studied by simulation. An already established multi-state model for non-small cell lung cancer is used to illustrate the adaptive procedure.

**Poster**

# On variance estimation for the one-sample log-rank test

**Moritz Fabian Danzer, Andreas Faldum, Rene Schmidt**
Institute of Biostatistics and Clinical Research, University of Münster, Germany

Time-to-event endpoints show an increasing popularity in phase II cancer trials. The standard statistical tool for such endpoints in one-armed trials is the one-sample log-rank test. It is widely known, that the asymptotic providing the correctness of this test does not come into effect to full extent for small sample sizes. There have already been some attempts to solve this problem. While some do not allow easy power and sample size calculations, others lack a clear theoretical motivation and require further considerations. The problem itself can partly be attributed to the dependence of the compensated counting process and its variance estimator. We provide a framework in which the variance estimator can be flexibly adopted to the present situation while maintaining its asymptotical properties. We exemplarily suggest a variance estimator which is uncorrelated to the compensated counting process. Furthermore, we provide sample size and power calculations for any approach fitting into our framework. Finally, we compare several methods via simulation studies and the hypothetical setup of a Phase II trial based on real world data.

**Contributed Talk**

# IDENTIFYING OPTIMIZED DECISION CRITERIA AND EXPERIMENTAL DESIGNS BY SIMULATING PRECLINICAL EXPERIMENTS IN SILICO

**Meggie Danziger**[1,2], **Ulrich Dirnagl**[1,2], **Ulf Tölch**[2]

[1]Charité – Universitätsmedizin Berlin, Germany; [2]BIH QUEST Center for Transforming Biomedical Research

Low statistical power in preclinical experiments has been repeatedly pointed out as a roadblock to successful replication and translation. To increase reproducibility of preclinical experiments under ethical and budget constraints, it is necessary to devise strategies that improve the efficiency of confirmatory studies.

To this end, we simulate two preclinical research trajectories from the exploratory stage to the results of a within-lab replication study based on empirical pre-study odds. In a first step, a decision is made based on exploratory data whether to continue to a replication. One trajectory (T1) employs the conventional significance threshold for this decision. The second trajectory (T2) uses a more lenient threshold based on an a priori determined smallest effect size of interest (SESOI). The sample size of a potential replication study is calculated via a standard power analysis using the initial exploratory effect size (T1) or using a SESOI (T2). The two trajectories are compared regarding the number of experiments proceeding to replication, number of animals tested, and positive predictive value (PPV).

Our simulations show that under the conventional significance threshold, only 32 percent of the initial exploratory experiments progress to the replication stage. Using the decision criterion based on a SESOI, 65 percent of initial studies proceed to replication. T1 results in the lowest number of animals needed for replication (n = 7 per group) but yields a PPV that is below pre-study odds. T2 increases PPV above pre-study odds while keeping sample size at a reasonably low number (n = 23 per group).

Our results reveal that current practice, represented by T1, impedes efforts to replicate preclinical experiments. Optimizing decision criteria and experimental design by employing easily applicable variations as shown in T2 keeps tested animal numbers low while generating more robust preclinical evidence that may ultimately benefit translation.

**Contributed Talk**

# Variable Importance Measures for Functional Gradient Descent Boosting Algorithm

**Zeyu Ding**
TU Dortmund, Germany

With the continuous growth of data dimensions and the improvement of computing power in contemporary statistics, the number of variables that can be included in a model is increasing significantly. Therefore, when designing a model, selecting truly informative variables from a bunch of messy variables and ranking them according to their importance becomes a core problem of statistics. Appropriate variable selection can reduce overfitting and leads to a more interpretable model. Traditional methods use the decomposition of R2, step-wise Akaike Information Criterion (AIC)-based variable selection, or regularization based on lasso and ridge. In ensemble algorithms, the variable importance is often calculated separately by the permutation methods. In this contribution, we propose two new stable and discriminating variable importance measures for the functional gradient descent boosting algorithm (FGDB). The first one calculates the l2 norm contribution of a variable while the second one calculates the risk reduction of the variables in every iteration. Our proposal is demonstrated in both simulation and real data examples. We show that the two new methods are more effective in automatically selecting those truly important variables in different data scenarios than the traditional selection frequency measures used in FGDB algorithm. This holds for both linear and non-linear models under different data scenarios.

Keywords: variable importance measures, variable selection, functional gradient boosting, component-wise regression, generalized additive models.

**Contributed Talk**

# CASANOVA: Permutation inference in factorial survival designs

**Marc Ditzhaus**[1], **Arnold Janssen**[2], **Markus Pauly**[1]
[1]TU Dortmund, Germany; [2]Heinrich-Heine-University Duesseldorf

In this talk, inference procedures for general factorial designs with time-to-event endpoints are presented. Similar to additive Aalen models, null hypotheses are formulated in terms of cumulative hazards. Deviations are measured in terms of quadratic forms in Nelson--Aalen-type integrals. Different to existing approaches, this allows to work without restrictive model assumptions as proportional hazards. In particular, crossing survival or hazard curves can be detected without a significant loss of power. For a distribution-free application of the method, a permutation strategy is suggested. The theoretical findings are complemented by an extensive simulation study and the discussion of a real data example.

**Contributed Talk**

# Which Test for Crossing Survival Curves? A User's Guide

**Ina Dormuth[1], Tiantian Liu[2,4], Jin Xu[2], Menggang Yu[3], Markus Pauly[1], Marc Ditzhaus[1]**

[1]TU Dortmund, Deutschland; [2]East China Normal University, China; [3]University of Wisconsin-Madison, USA; [4]Technion - Israel Institute of Technology, Israel

Knowledge transfer between statisticians developing new data analysis methods, and users is essential. This is especially true for clinical studies with time-to-event endpoints. One of the most common problems is the comparison of survival in two-armed trials. The log-rank test is still the gold standard for answering this question. However, in the case of non-proportional hazards, its informative value may decrease. In the meantime several extensions have been developed to solve this problem. Since non-proportional or even intersecting survival curves are common in oncology, e.g. in immunotherapy studies, it is important to identify the most appropriate methods and to draw attention to their existence. Therefore, it is our goal to simplify the choice of a test to detect differences in survival rate in case of crossings. To this end, we reviewed 1,400 recent oncological studies. Limiting our analysis to intersecting survival curves and non-significant log-rank tests for a sufficient number of observed events we reconstructed the data sets using a state-of-the-art reconstruction algorithm. To ensure reproductive quality, only publications with a published number of risk at multiple points in time, sufficient print quality, and a non-informative censoring pattern were included. After elimination of papers on the basis of the exclusion criteria mentioned above, we compared the p-values of the log-rank and the Peto-Peto test as references and compare them with nine different tests for non-proportional or even crossing hazards. It is shown that tests designed to detect crossing hazards are advantageous and provide guidance in choosing a reasonable alternative to the standard log-rank test. This is followed by a comprehensive simulation study and the generalization of one of the test methods to the multi-sample case.

**Bernd-Streitberg Laureate**

# Model selection characteristics when using MCP-Mod for dose-response gene expression data

**Julia Christin Duda**
TU Dortmund University, Germany

Classical approaches in clinical dose-finding trials rely on pairwise comparisons between doses and placebo. A methodological improvement is the MCP-Mod (Multiple Comparison Procedure and Modeling) approach, originally developed for Phase II trials. MCP-Mod combines multiple comparisons with modeling approaches in a multistage procedure. First, for a set of pre-specified candidate models, it is tested if any dose-response signal is present. Second, considering models with detected signal, either the best model is selected to fit the dose-response curve or model averaging is performed.

We extend the scope of application for MCP-Mod to in-vitro gene expression data and assess its characteristics regarding model selection for concentration gene expression curves. Precisely, we apply MCP-Mod on single genes of a high-dimensional gene expression data set, where human embryonic stem cells were exposed to eight concentration levels of the compound valproic acid (VPA). As candidate models we consider the sigmoid Emax (four-parameter log-logistic), linear, quadratic, Emax, exponential and beta model. Through simulations, we investigate the impact of omitting one or more models from the candidate model set to uncover possibly superfluous models and the precision and recall rates of selected models. Measured by the AIC, all models perform best for a considerable number of genes. For less noisy cases the popular sigmoid Emax model is frequently selected. For more noisy data, often simpler models like the linear model are selected, but mostly without relevant performance advantage compared to the second-best model. Also, the commonly used Emax model has an unexpected low performance.

**Contributed Talk**

# An intuitive time-dose-response model for cytotoxicity data with varying exposure times

**Julia Christin Duda, Jörg Rahnenführer**
TU Dortmund University, Germany

Modeling approaches for dose-response or concentration-response analyses are slowly becoming more popular in toxicological applications. For cytotoxicity assays, not only the concentration but also the exposure or incubation time of the compound administered to cells can be varied and might have influence on the response. A popular concentration-response model is the four-parameter log-logistic (4pLL) or, more specific and tailored to cytotoxicity data, the two-parameter log-logistic (2pLL) model. Both models, however, model the response based on the concentration only.

We propose a two-step procedure and a new time-concentration-response model for cytotoxicity data in which both concentration and exposure time are varied. The parameter of interest for the estimation is the EC50 value, i.e. the concentration at which half of the maximal effect is reached. The procedure consists of a testing step and a modeling step. In the testing step, a nested ANOVA test is performed to decide if the exposure time has an effect on the shape of the concentration-response curve. If no effect is identified then a classical 2pLL model is fitted. Otherwise, a new time-concentration-response model called td2pLL is fitted. In this model, we incorporate exposure time information into the 2pLL model by making the EC50 parameter dependent on the exposure time.

In simulation studies inspired by and based on a real data set, we compare the proposed procedure against various alternatives with respect to the precision of the estimation of the EC50 value. In all simulations, the new procedure provides estimates with higher or comparable precision, which demonstrates its universal applicability in corresponding toxicological experiments. In addition, we show that the use of optimal designs for cytotoxicity experiments further improves the EC50 estimates throughout all considered scenarios while reducing numerical problems. In order to facilitate the application in toxicological practice, the developed methods will be made available to practitioners via the R package td2pLL and a corresponding vignette that demonstrates the application on an example dataset.

**Contributed Talk**

# Reproducible bioinformatics workflows: A case study with software containers and interactive notebooks

**Anja Eggert, Pal O Westermark**
Leibniz Institute for Farm Animal Biology, Deutschland

We foster transparent and reproducible workflows in bioinformatics, which is challenging given their complexity. We developed a new statistical method in the field of circadian rhythmicity, which allows to rigorously determine whether measured quantities such as gene expressions are not rhythmic. Knowledge of no or at most weak rhythmicity may significantly simplify studies, aid detection of abolished rhythmicity, and facilitate selection of non-rhythmic reference genes or compounds, among other applications. We present our solution to this problem in the form of a precisely formulated mathematical statistic accompanied by a software called SON (Statistics Of Non-rhythmicity). The statistical method itself is implemented in the R package “HarmonicRegression”, available on the CRAN repository. However, the bioinformatics workflow is much larger than the statistical test. For instance, to ensure the applicability and validity of the statistical method, we simulated data sets of 20,000 gene expressions over two days, with a large range of parameter combinations (e.g. sampling interval, fraction of rhythmicity, amount of outliers, detection limit of rhythmicity, etc.). Here we describe and demonstrate the use of a Jupyter notebook to document, specify, and distribute our new statistical method and its application to both simulated and experimental data sets. Jupyter notebooks combine text documentation with dynamically editable and executable code and are an implementation of the concept of literate programming. Thus, parameters and code can be modified, allowing both verification of results, as well as instant experimentation by peer reviewers and other users of the science community. Our notebook runs inside a Docker software container, which mirrors the original software environment. This approach avoids the need to install any software and ensures complete long-term reproducibility of the workflow. This bioinformatics workflow allows full reproducibility of our computational work.

**Contributed Talk**

# Using Historical Data to Predict Health Outcomes – The Prediction Design

**Stella Erdmann, Manuel Feißt, Johannes Krisam, Meinhard Kieser**
Institute of Medical Biometry and Informatics, University of Heidelberg, Germany

The gold standard for the investigation of the efficacy of a new therapy is a randomized controlled trial (RCT). This is costly, time consuming and not always practicable (e.g. for lethal conditions with limited treatment possibilities) or even possible in a reasonable time frame (e.g. in rare diseases due to small sample sizes). At the same time, huge quantities of available control-condition data in analyzable format of former RCTs or real-world data (RWD), i.e., patient-level data gathered outside the conventional clinical trial setting, are neglected if not often completely ignored. To overcome these shortcomings, alternative study designs using data more efficiently would be desirable.

Assuming that the standard therapy and its mode of functioning is well known and large volumes of patient data exist, it is possible to set up a sound prediction model to determine the treatment effect of this standard therapy for future patients. When a new therapy is intended to be tested against the standard therapy, the vision would be to conduct a single-arm trial and to use the prediction model to determine the effect of the standard therapy on the outcome of interest of patients receiving the test treatment only, instead of setting up a two-arm trial for this comparison. While the advantages of using historical data to estimate the counterfactual are obvious (increased efficiency, lower cost, alleviating participants' fear of being on placebo), bias could be caused by confounding (e.g. by indication, severity, or prognosis) or a number of other data issues that could jeopardize the validity of the non-randomized comparison.

The aim is to investigate if and how such a design - the prediction design - may be used to provide information on treatment effects by leveraging existing infrastructure and data sources (historical data of RCTs and/or RWD). Therefore, we investigate under what assumptions a linear prediction model could be used to predict the counterfactual of patients precisely enough to construct a test for evaluating the treatment effect for normally distributed endpoints. In particular, it is investigated what amount of data is necessary (for the historical data and for the single arm trial to be conducted). Via simulation studies, it is examined how sensible the design acts towards violations of the assumptions. The results are compared to reasonable (conventional) benchmark scenarios, e.g., the setting of a single-arm study with pre-defined threshold or a setting, where a propensity score matching was performed.

**Contributed Talk**

# Adapting Variational Autoencoders for Realistic Synthetic Data with Skewed and Bimodal Distributions

**Kiana Farhadyar, Harald Binder**
Faculty of Medicine and Medical Center – University of Freiburg, Germany

**Background:**

Passing synthetic data instead of original data to the other researchers is an option when data protection restrictions exist. Such data should preserve the statistical relationships between the variables while protecting privacy. In recent years, deep generative models have allowed for significant progress in the field of synthetic data generation. In particular, variational autoencoders (VAEs) are a popular class of deep generative models. Standard VAEs are typically built around a latent space with a Gaussian distribution and this is a key challenge for VAEs when they encounter more complex data distributions like bimodal or skewed data.

**Methods:**

In this work, we propose a novel method for synthetic data generation that handles bimodal and skewed data as well, while keeping the overall VAE framework. Moreover, this method can generate synthetic data for datasets consisting of both continuous and binary variables. We apply two transformations to convert the data into a form that is more compliant with VAEs. First, we use Box-Cox transformations to transform the skewed distribution to something closer to a symmetric distribution. Then, dealing with potential bimodal data, we employ a power function sgn(x)|x|^p that can transform the data in a way that it has closer peaks and lighter tails. For the evaluation, we use a simulation design data, which is based on a large breast cancer study and The International Stroke Trial (IST) dataset as a real data example.

**Results:**

We show that the pre-transformations can make a considerable improvement in the utility of synthetic data for skewed and bimodal distributions. We investigate this in comparison with standard VAEs and a VAE with an autoregressive implicit quantile network approach (AIQN) and also Generative Adversarial Networks (GAN). We see that our method is the only method that can generate bimodality and the other methods typically generate unimodal distributions. For skewed data, these methods decrease the skewness of synthetic data and make the data closer to a symmetric distribution while our method produces similar skewness to original data and honors the value range of original data better.

**Conclusion:**

In conclusion, we developed a simple method, which adapts VAEs by transformations to handle skewed and bimodal data. Due to its simplicity, it is possible to combine it with many extensions of VAEs. Thus, it becomes feasible to generate high-quality synthetic clinical data for research under data protection constraints.

**Contributed Talk**

# Sampling designs for rare time-dependent exposures - A comparison of the nested exposure case-control design and exposure density sampling

**Jan Feifel[1], Maja von Cube[2], Martin Wolkewitz[2], Jan Beyersmann[1], Martin Schumacher[2]**

[1]Institute of Statistics, Ulm University, Germany; [2]Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center University of Freiburg, Germany

Hospital-acquired infections increase both morbidity and mortality of hospitalized patients. Researchers interested in the effect of these time-dependent infections on the length-of-hospital stay, as a measure of disease burden, face large cohorts with possibly rare exposures.

For large cohort studies with rare outcomes nested case-control designs are favorable due to the efficient use of limited resources. Here, nested case-control designs apply but do not lead to reduced sample sizes, because the outcome is not necessarily rare, but the exposure is. Recently, exposure density sampling (EDS)[1] and nested exposure case-control design (NECC) [2] have been proposed to sample for a rare time-dependent exposure in cohorts with a survival endpoint. The two designs differ in the time point of sampling.

Both designs enable efficient hazard ratio estimation by sampling all exposed individuals but only a small fraction of the unexposed ones. Moreover, they account for time-dependent exposure to avoid immortal time bias. We investigate and compare their performance using data of patients hospitalized in the neuro-intensive care unit at the Burdenko Neurosurgery Institute (NSI) in Moscow, Russia. The impact of different types of hospital-acquired infections with different prevalence on length-of-stay is considered. Additionally, inflation factors, a primary performance measure, are discussed. All presented methods will be compared to the gold-standard Cox model on the full cohort. We enhance both designs to allow for a competitive analysis of combined and competing endpoints. Additionally, these designs substantially reduce the amount of necessary information compared to the full cohort approach.

Both EDS and NECC are capable of analyzing time-to-event data by simultaneously accounting for rare time-dependent exposure and result in affordable sample sizes. EDS outperforms the NECC concerning efficiency and accuracy in most considered settings for combined endpoints. For competing risks, however, a tailored NECC shows more appealing results.

**References:**

[1] K. Ohneberg, J. Beyersmann and M. Schumacher (2019): Exposure density sampling: Dynamic matching with respect to a time-dependent exposure. Statistics in Medicine, 38(22):4390-4403.

[2] J. Feifel, M. Gebauer, M. Schumacher and J. Beyersmann (2020): Nested exposure case-control sampling: a sampling scheme to analyze rare time-dependent exposures. Lifetime Data Analysis, 26:21-44.

**Contributed Talk**

# Adaptive group sequential survival comparisons based on log-rank and pointwise test statistics

**Jannik Feld, Andreas Faldum, Rene Schmidt**
Institute of Biostatistics and Clinical Research, University of Münster

Whereas the theory of confirmatory adaptive designs is well understood for uncensored data, implementation of adaptive designs in the context of survival trials remains challenging. Commonly used adaptive survival tests are based on the independent increments structure of the log-rank statistic. These designs suffer the limitation that effectively only the interim log-rank statistic may be used for design modifications (such as data-dependent sample size recalculation). Alternative approaches based on the patient-wise separation principle have the common disadvantage that the test procedure may either neglect part of the observed survival data or tends to be conservative. Here, we instead propose an extension of the independent increments approach to adaptive survival tests. We present a confirmatory adaptive two-sample log-rank test of no difference in a survival analysis setting, where provision is made for interim decision making based on both the interim log-rank statistic and/or pointwise survival-rates, while avoiding aforementioned problems. The possibility to include pointwise survival-rates eases the clinical interpretation of interim decision making and is a straight forward choice for seamless phase II/III designs. We will show by simulation studies that the performance does not suffer using the pointwise survival-rates and exemplary consider application of the methodology to a two-sample log-rank test with binding futility criterion based on the observed short-term survival-rates and sample size recalculation based on conditional power. The methodology is motivated by the LOGGIC Europe Trial from pediatric oncology. Distributional properties are derived using martingale techniques in the large sample limit. Small sample properties are studied by simulation.

**Contributed Talk**

# Causal inference methods for small non-randomized studies: Methods and an application in COVID-19

**Sarah Friedrich, Tim Friede**
University Medical Center Göttingen, Germany

The usual development cycles are too slow for the development of vaccines, diagnostics and treatments in pandemics such as the ongoing SARS-CoV-2 pandemic. Given the pressure in such a situation, there is a risk that findings of early clinical trials are overinterpreted despite their limitations in terms of size and design. Motivated by a non-randomized open-label study investigating the efficacy of hydroxychloroquine in patients with COVID-19, we describe in a unified fashion various alternative approaches to the analysis of non-randomized studies. A widely used tool to reduce the impact of treatment-selection bias are propensity score (PS) methods and g-computation. Conditioning on the propensity score allows one to replicate the design of a randomized controlled trial, conditional on observed covariates. Moreover, doubly robust estimators provide additional advantages. Here, we investigate the properties of propensity score based methods including three variations of doubly robust estimators in small sample settings, typical for early trials, in a simulation study.

**Invited Talk**

# AI Models for Multi-Modal Data Integration

**Holger Fröhlich**
Fraunhofer Gesellschaft e.V.

Precision medicine aims for the delivery of the right treatment for the right patients. One important goal is therefore to identify molecular sub-types of diseases, which opens the opportunity for a better targeted therapy of disease in the future. In that context high throughput omics data has been extensively used. However, analysis of one type of omics data alone provides only a very limited view on a complex and systemic disease such as cancer. Correspondingly, parallel analysis of multiple omics data types is needed and employed more and more routinely. However, leveraging the full potential of multi-omics data requires statistical data fusion, which comes along with a number of unique challenges, including differences in data types (e.g. numerical vs discrete), scale, data quality and dimension (e.g. hundreds of thousands of SNPs vs few hundred miRNAs).

In the first part of my talk I will focus on Pathway based Multi-modal AutoEncoders (PathME) as one possible approach for multi-omics data integration. PathME relies on a multi-modal sparse denoising autoencoder architecture to embed multiple omics types that can be mapped to the same biological pathway. We show that sparse non-negative matrix factorization applied to such embeddings result into well discriminated disease subtypes in several cancer types, which show clinically distinct features. Moreover, each of these subtypes can be associated to subtype-specific pathways, and for each of these pathways it is possible to disentangle the influence of individual omics features, hence providing a rich interpretation.

Going one step further in the second part of my talk I will focus on Variational Autoencoder Modular Bayesian Networks (VAMBN) as merger of Bayesian Networks and Variational Autoencoders to model multiple data modalities (including clinical assessment scores), also in a longitudinal manner. I will specifically demonstrate the application of VAMBN for modeling entire clinical studies in Parkinson's Disease (PD). Since VAMBN is generative the model can be used to simulate synthetic patients, also under counterfactual scenarios (e.g. age shift by 20 years, modification of disease severity at baseline), which could facilitate the design of clinical studies, sharing of data under probabilistic privacy guarantees and eventually allowing for finding "patients-like-me" within a broader, virtually merged meta-cohort.

**Contributed Talk**

# Exploring missing patterns and missingness mechanisms in longitudinal patient-reported outcomes using data from a non-randomized controlled trial study

**Pimrapat Gebert[1,2,3], Daniel Schindel[1], Johann Frick[1], Liane Schenk[1], Ulrike Grittner[2,3]**

[1]Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Institute of Medical Sociology and Rehabilitation Science; [2]Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Institute of Biometry and Clinical Epidemiology; [3]Berlin Institute of Health (BIH), Anna-Louisa-Karsch-Str. 2, 10178 Berlin, Germany

Missing data mechanism plays an important role in the context of handling and analyzing data subject to missings. Longitudinal patient-reported outcome measures (PROMs) are usually far from complete, especially in seriously ill patients. To choose an appropriate strategy for handling missing data, most statistical approaches require knowledge about missingness patterns and assumptions about the type of missing data mechanism. We demonstrated how to explore the missingness patterns and mechanisms using PROMs data including global health status/QoL (GH/QoL) in the EORTC QLQ-C30, Patient reaction assessment (PRA-D), The Revised Illness Perception Questionnaire (IPQ-R), German modified version of the Autonomy Preference Index (API-DM), Decision Conflict Scale (DCS), and European health literacy survey (HLS-EU-Q6) from the Oncological Social Care Project (OSCAR) study. Linear random-effects pattern-mixture models were performed for identifying missing not at random (MNAR) for each pattern. We found that the missing data on the GH/QoL in the EORTC QLQ-C30 could be assumed as MNAR in missing data due to massive worsening of health status and death. However, there was no evidence of MNAR in any other PROMs measures. Although determining the true missing data mechanism is impossible, a pattern-mixture model can be useful in evaluating the effects of informative missingness in longitudinal PROMs.

**Contributed Talk**

# Coronary artery calcification in the middle-aged and elderly population of Denmark

**Oke Gerke[1,2], Jes Sanddal Lindholt[1,3], Barzan Haj Abdo[1], Axel Cosmus Pyndt Diederichsen[1,4]**

[1]Dept. of Clinical Research, University of Southern Denmark, DK; [2]Dept. of Nuclear Medicine, Odense University Hospital, DK; [3]Dept. of Cardiothoracic and Vascular Surgery, Odense University Hospital, DK; [4]Dept. of Cardiology, Odense University Hospital, DK

**Aims:**

Coronary artery calcification (CAC) measured on cardiac CT is an important risk marker for cardiovascular disease (CVD), and has been included in the prevention guidelines. The aim of this study was to describe CAC score reference values and to develop a free available CAC calculator in the middle-aged and elderly population. This work updates two previously published landmark studies on CAC score reference values, the American MESA study and the German HNR study [1,2]. Differences in curve-derivation compared to a recently published pooled analysis are discussed [3].

**Methods:**

17,252 participants from two population-based cardiac CT screening cohorts (DanRisk and DANCAVAS) were included [4,5]. The CAC score was measured as a part of s screening session. Positive CAC scores were log-transformed and nonparametrically regressed on age for each gender, and percentile curves were transposed according to proportions of zero CAC scores.

**Results:**

Men had higher CAC scores than women, and the prevalence and extend of CAC increased steadily with age. An online CAC calculator was developed, http://flscripts.dk/cacscore/. After entering sex, age and CAC score, the CAC score percentile and the coronary age are depicted including a figure with the specific CAC score and 25%, 50%, 75% and 90% percentiles. The specific CAC score can be compared to the entire background population or only those without prior CVD.

**Conclusion:**

This study provides modern population-based reference values of CAC scores in men and woman, and a freely accessible online CAC calculator. Physicians and patients are very familiar with blood pressure and lipids, but unfamiliar with CAC scores. Using the calculator makes it easy to see if a CAC value is low, moderate or high, when a physician in the future communicates and discusses a CAC score with a patient.

**References:**

[1] Schmermund A et al. Population-based assessment of subclinical coronary atherosclerosis using electron-beam computed tomography. Atherosclerosis 2006;185(1):177-182.

[2] McClelland RL et al. Distribution of coronary artery calcium by race, gender, and age: results from the Multi-Ethnic Study of Atherosclerosis (MESA). Circulation 2006;113(1):30-37.

[3] de Ronde MWJ et al. A pooled-analysis of age and sex based coronary artery calcium scores percentiles. J Cardiovasc Comput Tomogr. 2020;14(5):414-420.

[4] Diederichsen AC et al. Discrepancy between coronary artery calcium score and HeartScore in middle-aged Danes: the DanRisk study. Eur J Prev Cardiol 2012;19(3):558-564.

[5] Diederichsen AC et al. The Danish Cardiovascular Screening Trial (DANCAVAS): study protocol for a randomized controlled trial. Trials 2015;16:554.

**Invited Talk**

# On the Role of Historical Control Data in Preclinical Development

**Helena Geys**
Johnson & Johnson, Belgium

Historical control databases are established by many companies in order to be able to contextualize results from single studies against previous studies performed under similar conditions, to properly design studies and/or to come up with quality control instruments.

Typical preclinical experiments involve a study of a control group of untreated animals and groups of animals exposed to increasing doses. The ultimate aim is to test for a dose related trend in the response of interest. Usually one would focus on one particular experiment. However, since such experiments are conducted in genetically homogeneous animal strains, historical control data from previous similar experiments are sometimes used in interpreting results of a current study.

The use of historical control data in supporting inferences varies across different assays. For example, in genetic toxicology and safety pharmacology, a response may be considered positive in a specific experiment if the result is outside the distribution of the historical negative control data (95% control limits). Whereas, in carcinogenicity studies, historical control data are particularly useful in classifying tumors as rare or common and for evaluation of disparate findings in dual concurrent controls.

Historical control data are often used to carry out an informal equivalence test, whereby a New Molecular Entity (NME) is considered to be “safe” when the results from the treatment groups fall entirely within the negative control distribution.

In addition, formal statistical procedures have been proposed that allow to incorporate historical control data and to combine them with the current control group in tests trend identification.

Clearly historical control data are playing an important role in preclinical development as quality control and interpretation instrument. Yet, the issue of when and how to use historical control data is still not clear and subject to ongoing debate.

In this presentation we will highlight pros and cons and the important role a preclinical statistician can play in this.

**Invited Talk**

# Learning about personalised effects: transporting anonymized information from individuals to (meta-) analysis and back

**Els Goetghebeur**
Ghent University, Belgium

Evidence on `personalized' or stratified medicine draws information from subject-specific records on prognostic factors, (point) treatment and outcome from relevant population samples. A focus on treatment by covariate interactions makes such studies more data hungry than a typical trial focusing on a population average effect. Nationwide disease registers or individual patient meta-analysis may overcome sample size issues, but encounter new challenges, especially when targeting a risk or survival outcome. Between study heterogeneity comes as a curse and a blessing when aiming to transport treatment effects to new patient populations. It reveals sources of variation with specific roles in the transportation. For survival analysis special attention must be given to calendar time and internal time. A core set of covariates is needed for the stratified analysis, ideally measured with similar precision. A shared minimum follow-up time, and well understood censoring mechanisms are expected. Variation in baseline hazards under standard of care may reflect between study variation in diagnostic criteria, in populations, in unmeasured baseline covariates, in standard of care delivery and its impact.

When core set covariates are lacking for some studies or over certain stretches of time, a range of solutions may be considered. Different assumptions on missing covariates of survival models will affect treatment balancing IPW and direct standardization methods differentially. Alternatively, one may seek to link data from different sources to fill the gaps. It then pays to consider models with estimators that can be calculated from (iteratively) constructed summary statistics involving weighted averages over functions of the missing covariates. By thus avoiding the need for additional individually linked measures, one may open access to a range of existing covariates or biomarkers that can be merged while circumventing (time)consuming lengthy confidentiality agreements.

In light of the above we discuss pros and cons of various methods of standardizing effects to obtain transportable answers that are meaningfully compared between studies. We thus aim to provide relevant evidence on stratified interventions referring to several case studies.

**Invited Talk**

# Marginalized Frailty-Based Illness-Death Model: Application to the UK-Biobank Survival Data

**Malka Gorfine**
Tel Aviv University, Israel

The UK Biobank is a large-scale health resource comprising genetic, environmental, and medical information on approximately 500,000 volunteer participants in the United Kingdom, recruited at ages 40–69 during the years 2006–2010. The project monitors the health and well-being of its participants. This work demonstrates how these data can be used to yield the building blocks for an interpretable risk-prediction model, in a semiparametric fashion, based on known genetic and environmental risk factors of various chronic diseases, such as colorectal cancer. An illness-death model is adopted, which inherently is a semi-competing risks model, since death can censor the disease, but not vice versa. Using a shared-frailty approach to account for the dependence between time to disease diagnosis and time to death, we provide a new illness-death model that assumes Cox models for the marginal hazard functions. The recruitment procedure used in this study introduces delayed entry to the data. An additional challenge arising from the recruitment procedure is that information coming from both prevalent and incident cases must be aggregated. Lastly, we do not observe any deaths prior to the minimal recruitment age, 40. In this work, we provide an estimation procedure for our new illness-death model that overcomes all the above challenges.

**Contributed Talk**

# Modelling acute myeloid leukemia: Closing the gap between model parameters and individual clinical patient data

**Dennis Görlich**
Institute of Biostatistics and Clinical Research, University Münster, Germany

In this contribution, we will illustrate and discuss our approach to fit a mechanistic mathematical model of acute myeloid leukemia (AML) to individual patient data, leading to personalized model parameter estimates.

We use a previously published model (Banck and Görlich, 2019) that describes the healthy hematopoiesis and the leukemia dynamics. Here, we consider a situation where the healthy hematopoiesis is calibrated to a population average and personalized leukemia parameters (self renewal, proliferation, and treatment intensity) needs to be estimated.

To link the mathematical model to clinical data model predictions needs to be aligned to observable clinical outcome measures. In AML research, blast load, complete remission, and survival are typically considered. Based on the model's properties, especially the capability to predict the considered outcomes, blast load turned out to be well suited for the model fitting process.

We formulated an optimization problem to estimate personalized model parameters based on the comparison between observed and predicted blast load (cf. Görlich, 2021).

A grid search was performed to evaluate the fitness landscape of the optimization problem. The grid search approach showed that, depending on the patient's individual blast course, noisy fitness landscapes can occur. In these cases, a gradient-descent algorithm will usually perform poorly. This problem can be overcome by application of e.g. the differential evolution algorithm (Price et al., 2006). The estimated personalized leukemia parameters can be further correlated to observed clinical data. A preliminary analysis showed promising results.

Finally, the application of mechanistic mathematical models in combination with personalized model fitting seems to be a promising approach within clinical research.

### References

[1] Dennis Görlich (accepted). Fitting Personalized Mechanistic Mathematical Models of Acute Myeloid Leukaemia to Clinical Patient Data. Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies, Volume 3: BIOINFORMATICS 2021

[2] Jan C. Banck and Dennis Görlich (2019). In-silico comparison of two induction regimens (7 + 3 vs 7 + 3 plus additional bone marrow evaluation) in acute myeloid leukemia treatment. BMC Systems Biology, 13(1):18.

[3] Kenneth V. Price, Rainer M. Storn and Jouni A. Lampinen (2006). Differential Evolution - A Practical Approach to Global Optimization. Berlin Heidelberg: Springer-Verlag.

**Contributed Talk**

# Mixed-effects ANCOVA for estimating the difference in population mean parameters in case of nonlinearly related data

**Ricarda Graf**
University of Göttingen, Germany

Repeated measures data can be found in many fields. The two types of variation characteristic for this type of data - referred to as within-subject and between-subject variation – are accounted for by linear and nonlinear mixed-effects models. ANOVA-type models are sometimes applied for comparison of population means despite a nonlinear relationship in the data. Accurate parameter estimation through more appropriate nonlinear-mixed effects (NLME) models, such as for sigmoidal curves, might be hampered due to insufficient data near the asymptotes, the choice of starting values for the iterative optimization algorithms used given the lack of closed-form expressions of the likelihood or due to convergence problems of these algorithms.

The main objective of this thesis is to compare the performance of a one-way mixed-effects ANCOVA and a NLME three-parameter logistic regression model with respect to the accuracy in estimating the difference in population means. Data from a clinical trial1, in which the difference in mean blood pressure (BP50) between two groups was estimated by repeated-measures ANOVA, served as a reference for data simulation. A third simplifying method, used in toxicity studies[2], was additionally included. It considers the two measurements per subject lying immediately below and above mean half maximal response (E_max). Population means are obtained by considering the intersections of the horizontal line represented by half E_max and the line derived from connecting the two data points per subject and group. A simulation study with two scenarios was conducted to compare bias, coverage rates and empirical SE of the three methods when estimating the difference in BP50 for purpose of identification of the disadvantages by using the simpler linear instead of the nonlinear model. In the first scenario, the true individual blood pressure ranges were considered, while in the second scenario, measurements at characteristic points of the sigmoidal curves were considered, regardless of the true measurement ranges, in order to obtain a more distinct nonlinear relationship.

The estimates of the mixed-effects ANCOVA model were more biased but also more precise compared with the NLME model. The ANCOVA method could not detect the difference in BP50 in the second scenario anymore. The results of the third method did not seem reliable since its estimates did on average even reverse the direction of the true parameter.

NLME models should be preferred for data with a known nonlinear relationship if the available data allows it. Convergence problems can be overcome by using a Bayesian approach.

**Poster**

# Quantification of severity of alcohol harms from others' drinking items using item response theory (IRT)

**Ulrike Grittner[1], Kim Bloomfield[1,2,3,4], Sandra Kuntsche[5], Sarah Callinan[5], Oliver Stanesby[5], Gerhard Gmel[6,7,8,9]**

[1]Institute of Biometry and Clinical Epidemiology, Charité - Universitätsmedizin Berlin, Germany; [2]Centre for Alcohol and Drug Research, Aarhus University, Denmark; [3]Research Unit for Health Promotion, University of Southern Denmark, Denmark; [4]Alcohol Research Group, Emeryville, CA, USA; [5]Centre for Alcohol Policy Research, La Trobe University, Melbourne, Australia; [6]Alcohol Treatment Centre, Lausanne University Hospital CHUV, Lausanne, Switzerland; [7]Addiction Switzerland, Research Department, Lausanne, Switzerland; [8]Centre for Addiction and Mental Health, Institute for Mental Health Policy Research, Toronto, Ontario, Canada; [9]University of the West of England, Faculty of Health and Applied Science, Bristol, United Kingdom

**Background:**

Others' heavy drinking might negatively affect quality of life, mental and physical health as well as work and family situation. However, until now there is little known about which of these experiences is seen as most or least harmful, and who is most affected.

**Methods:**

Data stem from large population-based surveys from 10 countries of the GENAHTO project (GENAHTO: Gender & Alcohol's Harms to Others, www.genahto.org ). Questions about harms from others' heavy drinking concern verbal and physical harm, damage of clothing, belongings or properties, traffic accidents, harassment, threatening behaviour, family problems, problems with friends, problems at work, and financial problems. We used item response theory (IRT) methods (two-parameter logistic (2PL) model) to allow for scaling of the aforementioned items for each country separately. To acknowledge culturally-related sensibilities to experiences of harms in different countries, we also used differential item functioning (DIF). This resulted in country-wise standardised person-based parameters for each individual of each country indicating a quantified measure of load of AHTO. In multiple linear mixed models (random intercept for country) we analysed how load of AHTO was related to sex, age, own drinking and education.

**Results:**

Younger age, female sex and higher level of own drinking were related to a higher load of AHTO. However, interaction of age and own drinking indicated that only for younger age did own drinking level play a role.

**Conclusions:**

Using IRT, we were able to evaluate differing grades of severity in the experiences of harm from others' heavy drinking.

**Bernd-Streitberg Laureate**

# Temporal Dynamics in Generative Models

**Maren Hackenberg, Harald Binder**
Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Germany

**Abstract**

Uncovering underlying development patterns in longitudinal biomedical data is a first step towards understanding disease processes, but is complicated by the sparse time grid and individual-specific development patterns that often characterize such data. In epidemiological cohort studies and clinical registries, we are facing the question of what can be learned from the data in an early phase of the study, when only a baseline characterization and one follow-up measurement are available. Specifically, we considered a data scenario where an extensive characterisation is available at a baseline time point for each individual, but only a smaller subset of variables is measured again at an individually differing second time point, resulting in a very sparse (only two time points) and irregular time grid.

Inspired by recent advances that allow to combine deep learning with dynamic modeling, we employed a generative deep learning model to capture individual dynamics in a low-dimensional latent representation as solutions of ordinary differential equations (ODEs). Here, the variables measured only at baseline are used to infer individual-specific ODE parameters.

Additionally, we enriched the information of each individual by linking groups of individuals with similar underlying trajectories, which then serve as proxy information on the common temporal dynamics. Irregular spacing in time can thus be used to gain more information on individual dynamics by leveraging individuals' similarity.

Using simulated data, we showed that the model can recover individual trajectories from linear and non-linear ODE systems with two and four unknown parameters and infer groups of individuals with similar trajectories. The results illustrate that dynamic deep learning approaches can be adapted to such small data settings to provide an individual-level understanding of the dynamics governing individuals' developments.

**Contributed Talk**

# Using Differentiable Programming for Flexible Statistical Modeling

**Maren Hackenberg[1], Marlon Grodd[1], Clemens Kreutz[1], Martina Fischer[2], Janina Esins[2], Linus Grabenhenrich[2], Christian Karagiannidis[3], Harald Binder[1]**

[1]Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Germany; [2]Robert Koch Institute, Berlin, Germany; [3]Department of Pneumology and Critical Care Medicine, Cologne-Merheim Hospital, ARDS and ECMO Center, Kliniken der Stadt Köln, Witten/Herdecke University Hospital, Cologne, Germany

Differentiable programming has recently received much interest as a paradigm that facilitates taking gradients of computer programs. While the corresponding flexible gradient-based optimization approaches so far have been used predominantly for deep learning or enriching the latter with modeling components, we want to demonstrate that they can also be useful for statistical modeling per se, e.g., for quick prototyping when classical maximum likelihood approaches are challenging or not feasible.

In an application from a COVID-19 setting, we utilize differentiable programming to quickly build and optimize a flexible prediction model adapted to the data quality challenges at hand. Specifically, we develop a regression model, inspired by delay differential equations, that can bridge temporal gaps of observations in the central German registry of COVID-19 intensive care cases for predicting future demand. With this exemplary modeling challenge, we illustrate how differentiable programming can enable simple gradient-based optimization of the model by automatic differentiation. This allowed us to quickly prototype a model under time pressure that outperforms simpler benchmark models.

We thus exemplify the potential of differentiable programming also outside deep learning applications, to provide more options for flexible applied statistical modeling.

**Poster**

# Sample Size in Bioequivalence Cross-Over Trials with Balanced Incomplete Block Design

**Lina Hahn[1], Gerhard Nehmiz[2], Jan Beyersmann[1], Salome Mack[2]**

[1]Universität Ulm, Institut für Statistik, Deutschland; [2]Boehringer Ingelheim Pharma GmbH&Co. KG, Biberach

In cross-over trials, all subjects receiving the same sequence of treatments form one sequence group, so there are s sequence groups of equal size n/s. If, due to limitations, the number of periods (p) is smaller than the number of treatments (t), we have an Incomplete Block Design. If the allocation of treatments to sequence groups is balanced, it is a Balanced Incomplete Block Design (BIBD).

Necessary conditions are: If r is the number of different sequences each treatment appears in, and if lambda is the number of different sequences in which each treatment pair occurs, balance implies r = p * s / t and lambda = r * (p-1) / (t-1). BIBDs exist for any t and p but can become large (Finney 1963). We investigate two examples with reasonable size, an internal one and the example of Senn (1997).

In medical trials, furthermore, period effects are likely, and in the BIBD the allocation of treatments to periods has also to be balanced, so that treatment contrasts can be estimated unbiasedly. Sufficient is that s is an integer multiple of t, or equivalently r is an integer multiple of p (Hartley 1953). Our two examples fulfil this.

Let the linear model for the measurements y be as usual with i.i.d. random subject effects and fixed terms for treatment, period and sequence group (absorbed by subject effect). All treatment contrasts can then be estimated in an unbiased manner, and if all error terms are i.i.d. N(0,sigma_e^2) and the subject effects are independent from these, the variance of the contrasts can be estimated as well. While in a complete cross-over the contrast variance is 2*sigma_e^2, it is in a BIBD generically b_k*sigma_e^2 where b_k is the “design factor”. We obtain b_k = (2*p*s) / (lambda*t).

Bioequivalence is investigated through the two one-sided tests procedure (TOST) for a treatment contrast (Schuirmann 1987). We investigate the power of the TOST in the two examples, considering the t distribution (Shen 2015, Labes 2020) and comparing it with a previous normal approximation which induces slight underpowerment.

**Contributed Talk**

# Academia-industry collaborations in biostatistics – It is not about the whether, just about the how

**Lisa Hampson**[1], **Frank Fleischer**[2]

[1]Advanced Methodology & Data Science, Novartis Pharma AG, Switzerland; [2]Biostatistics & Data Sciences, Boehringer Ingelheim Pharma, Germany

Methodological collaborations between academia and the pharmaceutical industry can have several benefits for both parties. In addition to the development and application of new statistical methods, there is also the education and recruitment of the next generation of biostatisticians and data scientists. In this presentation, we begin by reflecting on the key components (and maybe some pitfalls) of an academia-industry collaboration. We consider the different models that these collaborations can follow, ranging from co-supervision of student projects to collaborations between institutions. We will use several examples to illustrate the various models and their direct impact on statistical methodology and the business. Topics covered are diverse and range from data science to innovative clinical trial design. We conclude by looking to the future and provide an overview of emerging methodological questions in the pharmaceutical industry which we think are ripe for future academia-industry partnerships.

**Contributed Talk**

# Effect of missing values in multi-environmental trials on variance component estimates

**Jens Hartung, Hans-Peter Piepho**
University of Hohenheim, Germany

A common task in the analysis of multi-environmental trials (MET) by linear mixed models (LMM) is the estimation of variance components (VCs). Most often, MET data are imbalanced, e.g., due to selection. The imbalance mechanism can be missing completely at random (MCAR), missing at random (MAR) or missing not at random (MNAR). If the missing-data pattern in MET is not MNAR, likelihood-based methods are the preferred methods for analysis as they can account for selection. Likelihood-based methods used to estimate VCs in LMM have the property that all VC estimates are constrained to be non-negative and thus the estimators are generally biased. Therefore, there are two potential causes of bias in MET analysis: a MNAR data pattern and the small-sample properties of likelihood-based estimators. The current study tries to distinguish between both possible sources of bias. A simulation study with MET data typical for cultivar evaluation trials was conducted. The missing data pattern and size of VCs was varied. The results showed that for the simulated MET, VC estimates from likelihood-based methods are mainly biased due to the small-sample properties of likelihood-based methods for a small ratio of genotype variance to error variance.

**Invited Talk**

# The Statistical Assessment of Replication Success

**Leonhard Held**
Epidemiology, Biostatistics and Prevention Institute (EBPI) and Center for Reproducible Science (CRS), University of Zurich

Replicability of research findings is crucial to the credibility of all empirical domains of science. However, there is no established standard how to assess replication success and in practice many different approaches are used. Statistical significance of both the original and replication study is known as the two-trials rule in drug regulation but does not take the corresponding effect sizes into account.

We compare the two-trials rule with the sceptical p-value (Held, 2020), an attractive compromise between hypothesis testing and estimation. This approach penalizes shrinkage of the replication effect estimate compared to the original one, while ensuring that both are also statistically significant to some extent. We describe a recalibration of the procedure as proposed in Held et al (2020), the golden level. The golden level guarantees that borderline significant original studies can only be replicated successfully if the replication effect estimate is larger than the original one. The recalibrated sceptical p-value offers uniform gains in project power compared to the two-trials rule and controls the Type-I error rate except for very small replication sample sizes. An application to data from four large replication projects shows that the new approach leads to more appropriate inferences, as it penalizes shrinkage of the replication estimate compared to the original one, while ensuring that both effect estimates are sufficiently convincing on their own. Finally we describe how the approach can also be used to design the replication study based on specification of the minimum relative effect size to achieve replication success.

**References:**

[1] Held, Leonhard (2020) A new standard for the analysis and design of replication studies (with discussion). Journal of the Royal Statistical Society, Series A, 183:431–469.

[2] Held, Leonhard and Micheloud, Charlotte and Pawel, Samuel (2020). The assessment of replication success based on relative effect size. https://arxiv.org/abs/2009.07782

**Contributed Talk**

# Control of the population-wise error rate in group sequential trials with multiple populations

**Charlie Hillner, Werner Brannath**
Competence Center for Clinical Trials, Germany

In precision medicine one is often interested in clinical trials that investigate the efficacy of treatments that are targeted to specific sub-populations defined by genetic and/or clinical biomarkers. When testing hypotheses in multiple populations multiplicity adjustments are needed. First, we propose a new multiple type I error criterion for clinical trials with multiple intersecting populations, which is based on the observation that not all type I errors are relevant to all patients in the overall population. If the sub-populations are disjoint, no adjustment for multiplicity appears necessary, since a claim in one sub-population does not affect patients in the other ones. For intersecting sub-populations we suggest to control the probability that a randomly selected patient will be exposed to an inefficient treatment, which is an average multiple type I error. We propose group sequential designs that control the PWER where possibly multiple treatments are investigated in multiple populations. To this end, an error spending approach that ensures PWER-control is introduced. We exemplify this approach for a setting of two intersecting sub-populations and discuss how the number of different treatments to be tested in each sub-population affects the critical boundaries needed for PWER-control. Lastly, we apply this error spending approach to a group sequential design example from Magnusson & Turnbull (2013), where the efficacy of one treatment is to be tested after a certain sub-population that is likely to benefit from the treatment is found. We compare our PWER-controlling method with their FWER-controlling method in terms of critical boundaries and the resulting rejection probabilities and expected information.

**References:**

[1] Magnusson, B.P. and Turnbull, B.W. (2013), Group sequential enrichment design incorporating subgroup selection. Statist. Med., 32: 2695-2714. https://doi.org/10.1002/sim.5738

**Public Lecture**

# Statistik in Zeiten von Corona - Der komplexe Weg zur Zulassung eines Impfstoffes

**Benjamin Hofner**
Paul-Ehrlich-Institut, Bundesinstitut für Impfstoffe und biomedizinische Arzneimittel, Langen

Die schnelle Entwicklung eines wirksamen und sicheren Impfstoffes während einer Pandemie ist eine unglaubliche Herausforderung. Wie man im vergangenen Jahr in den Hauptnachrichten verfolgen konnte besteht diese Herausforderung nicht nur in der Entwicklung des Impfstoffes im Labor sondern auch in der darauf folgenden Testung und Zulassung des Impfstoffes.

Dieser Vortrag beleuchtet die klinische Entwicklung, in der letzten und entscheidenden Phase 3 Studie. Exemplarisch geht er dabei auf die Studien der zwei bereits zugelassenen Impfstoffe von BioNTech/Pfizer und Moderna ein. Er stellt die zentrale Rolle der statistischen Aspekte (z.B. Studiendesign, Fallzahlplanung, Studienpopulation, Endpunkte) sowohl bei der Planung als auch bei der Auswertung der Studien heraus. Statistische Konzepte und deren Notwendigkeit in Impfstoffstudien sollen dabei auch für interessierte Laien verständlich erklärt und motiviert werden. Außerdem werden regulatorische Abläufe die zur Zulassung der Impfstoffe führen aufgezeigt.

**Contributed Talk**

# A Web-Application to determine statistical optimal designs for dose-response trials, especially with interactions.

**Tim Holland-Letz, Annette Kopp-Schneider**
German Cancer Research Center DKFZ, Germany

Statistical optimal design theory is well developed, but almost never used in practical applications in fields such as toxicology. For the area of dose response trials we therefore present an R-shiny based web application which calculates D-optimal designs for the most commonly fitted dose response functions, namely the log-logistic and the Weibull function. In this context, the application also generates a graphical representation of the design space (a “design heatmap”). Furthermore, the application allows checking the efficiencies of user specified designs. In addition, uncertainty in regard to the assumptions about the true parameters can be included in the form of average optimal designs. Thus, the user can find a design which is a compromise between rigid optimality and more practical designs which also incorporate specific preferences and technical requirements.

Finally, the app can also be used to compute designs for substance interaction trials between two substances combined in a ray design setup, including an a-priori estimate for the parameters of the combination to be expected under the (Loewe-) additivity assumption.

**Invited Talk**

# Machine Learning in Biometry

**Chris Holmes**
University Of Oxford, Vereinigtes Königreich

Machine learning (ML) and artificial intelligence (AI) have had a major impact across many disciplines including biometrics. In the first half of this talk we will review some of the characteristics of ML that make for successful applications and also those features that present challenges, in particular around robustness and reproducibility. Relatively speaking, ML is mainly concerned with prediction while the majority of biometric analyses are focussed on inference. In the second half of the talk we will review the prediction-inference dichotomy and explore, from a Bayesian perspective, the theoretical foundations on how modern ML predictive models can be utilised for inference.

**Bernd-Streitberg Laureate**

# Internal validation for descriptive clustering of gene expression data

**Anastasiia Holovchak**
LMU Munich, Germany

**Abstract**

Cluster algorithms are often used to analyse gene expression data to partition the genes into homogenous groups based on their expression levels across patients. In practice, one is confronted with a large variety of clustering algorithms and it is often unclear which should be selected. A common procedure consists of testing different algorithms with several input parameters and evaluating them with appropriate internal cluster validation indices. However, it is again unclear which of these indices should be selected.

In this work, I conduct a study that investigates the stability of four internal cluster validation indices (Calinski-Harabasz index, Davies-Bouldin index, Dunn Index, and Average Silhouette Width criterion), in particular their ability to identify clusterings that replicate on independent test data. For the purpose of this study, an example gene expression data set is repeatedly split into a training and a test data set. Several commonly used clustering algorithms such as K-means, agglomerative clustering algorithms (Single Linkage, Complete Linkage, and Average Linkage), and spectral clustering algorithm are applied to the training data. The resulting clusterings are assessed using the four internal validation indices under consideration. The cluster methods are then applied to the test data and the similarity between the index values for the clusterings on the training and on the test data is assessed. I analyse whether the cluster algorithms and input parameters that are indicated as the best choices by the internal validation indices on the training data are also the best choices on the test data. Moreover, the internal validation indices are used to choose the best clustering on the training data and the stability of this selection process is investigated by applying the selected algorithm/parameter setting on the test data (as measured through the adjusted Rand index).

The results may guide the selection of appropriate indices in the considered context of gene expression data. For example, in this study the Dunn index yields very unstable results in terms of the selection of the best input parameter, which can be seen as an inconvenience. In conclusion, the investigated internal cluster validation indices show very different behaviours and one should not put much confidence in a single validation index unless there is evidence - from the literature or from own investigations such as the one presented in this thesis – that it yields meaningful replicable results in the considered context.

**Contributed Talk**

# Interaction forests: Identifying and exploiting influential quantitative and qualitative interaction effects

**Roman Hornung**
University of Munich, Germany

Even though interaction effects are omnipresent in biomedical data and play a particularly prominent role in genetics, they are given little attention in analysis, in particular in prediction modelling. Identifying influential interaction effects is valuable, both, because they allow important insights into the interplay between the covariates and because these effects can be used to improve the prediction performance of automatic prediction rules.

Random forest is one of the most popular machine learning methods and known for its ability to capture complex non-linear dependencies between the covariates and the outcome. A key feature of random forest is that it allows to rank the considered covariates with respect to their contribution to prediction using various variable importance measures.

We developed ‘interaction forest’, a variation of random forest for categorical, metric, and survival outcomes that explicitly considers several types of interaction effects in the splitting performed by the trees constituting the forest. The new ‘effect importance measure (EIM)’ associated with interaction forest allows to rank the interaction effects between the covariate pairs with respect to their importance for prediction in addition to ranking the univariable effects of the covariates in this respect. Using EIM, separate importance value lists for univariable effects, quantitative interaction effects, and qualitative interaction effects are provided. In a real data study using 220 publicly available data sets it is seen that the prediction performance of interaction forest is statistically significantly better than that of random forest and competing random forest variants that, as does interaction forest, use multivariable splitting. Moreover, a simulation study suggests that EIM allows to identify consistently the relevant quantitative and qualitative interaction effects in datasets. Here, the rankings obtained from the EIM value lists for quantitative interaction effects on the one hand and qualitative interaction effects on the other are confirmed to be specific for each of these two types of interaction effects. These results indicate that interaction forest is a suitable tool for identifying and making use of relevant interaction effects in prediction modelling.

**Contributed Talk**

# Sample size re-estimation based on the prevalence in a randomized test-treatment study

**Amra Hot, Antonia Zapf**
Institute of Medical Biometry and Epidemiology, University Medical Center Hamburg-Eppendorf, Hamburg

Patient benefit should be the primary criterion in evaluating diagnostic tests. If a new test has shown sufficient accuracy, its application in clinical practice should yield to a patient benefit. Randomized test-treatment studies are needed to assess the clinical utility of a diagnostic test as part of a broader management regimen in which test-treatment strategies are compared in terms of their impact on patient relevant outcomes [1]. Due to their increased complexity compared to common intervention trials the implementation of such studies poses practical challenges which might affect the validity of the study. One important aspect is the sample size determination. It is a special feature of these designs that they combine information on the disease prevalence and accuracy of the diagnostic tests, i.e. sensitivity and specificity of the investigated tests, with assumptions on the expected treatment effect. Due to the lack of empirical information or uncertainty regarding these parameters sample size consideration will always be based on a rather weak foundation, thus leading to an over- or underpowered trial. Therefore, it is reasonable to consider adaptations in earlier phases of the trial based on a pre-specified interim analysis in order to solve this problem. A blinded sample size re-estimation based on the disease prevalence in a randomized test-treatment study was performed as part of a simulation study. The type I error, the empirical overall power as well as the bias of the estimated prevalence are assessed and presented.

### References

[1] J. G. Lijmer, P.M. Bossuyt. Diagnostic testing and prognosis: the randomized controlled trial in test evaluation research. In: The evidence base of clinical diagnosis. Blackwell Oxford, 2009, 63-82.

**Contributed Talk**

# Statistical MODEling of Additive Time Effects in Survival Analysis

**Annika Hoyer**[1], **Oliver Kuß**[2]
[1]Department of Statistics, Ludwig-Maximilians-University Munich, Germany; [2]Institute for Biometrics and Epidemiology, German Diabetes Center, Leibniz Institute for Diabetes Research at Heinrich-Heine-University Duesseldorf, Germany

In survival analysis, there have been various efforts to model intervention or exposure effects on an additive rather than on a hazard, odds or accelerated life scale. Though it might be intuitively clear that additive effects can be easier understood, there is also evidence from randomized trials that this is indeed the case: treatment benefits are easier understood if communicated as postponement of an adverse event [1]. In clinical practice, physicians and patients tend to interpret an additive effect on the time scale as a gain in life expectancy which is added as additional time to the end of life [2]. However, as the gain in life expectancy is, from a statistical point of view, an integral, this is not a precise interpretation. As an easier interpretable alternative we propose to model the increasing „life span“ [3] and to examine the corresponding densities instead of the survival functions. Focussing on the respective modes, the difference of them describes a change in life span, especially the shifting of the most probable event time. Therefore, it seems reasonable to model differences in life time in terms of mode differences instead of differences in expected times. To this task, we propose mode regression models (which we write “Statistical MODEls” to emphasize that the modes are modelled) based on parametric distributions (Gompertz, Weibull and log-normal). We illustrate our MODEls by an example from a randomized controlled trial on efficacy of a new glucose-lowering drug for the treatment of type 2 diabetes.

**References:**

[1] Dahl R, Gyrd-Hansen D, Kristiansen IS, et al. Can postponement of an adverse outcome be used to present risk reductions to a lay audience? A population survey. BMC Med Inform Decis Mak 2007; 7:8

[2] Detsky AS, Redelmeier DA. Measuring health outcomes-putting gains into perspective. N Engl J Med 1998; 339:402-404

[3] Naimark D, Naglie G, Detsky AS. The meaning of life expectancy: what is a clinically significant gain? J Gen Intern Med 1994; 9:702-707

**Contributed Talk**

# A comparison of different statistical strategies for the analysis of data in reproductive toxicology involving historical negative controls

**Bernd-Wolfgang Igl, Monika Brüning, Bernd Baier**
Boehringer Ingelheim Pharma GmbH & Co. KG, Germany

A fundamental requirement of regulatory bodies for the development of new pharmaceuticals is to perform nonclinical developmental and reproductive toxicology (DART) studies to reveal any possible effect of the test item on mammalian reproduction. Usually DART studies are performed in rats and a further (non-rodent) species and aim to support human clinical trials and market access. General recommendations are given in ICH Guideline S5 allowing various phase-dependent designs for a huge number of parameters. The statistical evaluation of DART data is quite multifaceted due to more or less complex correlation structures between mother and offspring, e.g. maternal weight development, fetus weight, ossification status and number of littermates all in dependence of different test item doses.

Initially, we will sketch a Scrum inspired project that was set-up as a cooperation between Boehringer Ingelheim's Reproductive Toxicology and Non-Clinical Statistics groups. Then, we will describe the particular role and relevance of historical control data in reproductive toxicology. This will be followed by a presentation of common statistical models and some related open problems. Finally, we will give some simulation-based results on statistical power and sample size for the detection of certain events in DART studies.

**Contributed Talk**

# Correcting for bias due to misclassification in dietary patterns using 24 hour dietary recall data

**Timm Intemann**[1], **Iris Pigeot**[1,2]

[1]Leibniz Institute for Prevention Research and Epidemiology - BIPS, Germany; [2]Institute of Statistics, Faculty of Mathematics and Computer Science, University of Bremen, Germany

The development of statistical methods for nutritional epidemiology is a challenge, as nutritional data are usually multidimensional and error-prone. Analysing dietary data requires an appropriate method taking into account both multidimensionality and measurement error, but measurement error is often ignored when such data is analysed (1). For example, associations between dietary patterns and health outcomes are commonly investigated by first applying cluster analysis algorithms to derive dietary patterns and then fitting a regression model to estimate the associations. In such a naïve approach, errors in the underlying continuous dietary variables lead to misclassified dietary patterns and to biased effect estimates. To reduce this bias, we developed three correction algorithms for data assessed with a 24 hour dietary recall (24HDR), which has become the preferred dietary assessment tool in large epidemiological studies.

The newly developed correction algorithms combine the measurement error correction methods regression calibration (RC), simulation extrapolation (SIMEX) and multiple imputation (MI) with the cluster methods k-means cluster algorithm and the Gaussian mixture model. These new algorithms are based on univariate correction methods for Box-Cox transformed data (2) and consider the measurement error structure of 24HDR data. They consist mainly of the following three stages: (i) estimation of usual intakes, (ii) deriving patterns based on usual intakes and (iii) estimation of the association between these patterns and an outcome.

We apply the correction algorithms to real data from the IDEFICS/I.Family cohort to estimate the association between meal timing patterns and a marker for the long-term blood sugar level (HbA1c) in European children. Furthermore, we use the fitted parameters from this analysis to mimic the real cohort data in a simulation study. In this simulation study, we consider continuous and binary outcomes in different scenarios and compare the performance of the proposed correction algorithms and the naïve approach with respect to absolute, maximum and relative bias.

Simulation results show that the correction algorithms based on RC and MI perform better than the naïve and the SIMEX-based algorithms. Furthermore, the MI-based approach, which can use outcome information in the error model, is superior to the RC-based approach in most scenarios.

**References**

[1] Shaw, P. et al. (2018). Epidemiologic analyses with error-prone exposures: Review of current practice and recommendations. Ann Epidemiol 28, 821-828.

[2] Intemann, T. et al. (2019). SIMEX for correction of dietary exposure effects with Box-Cox transformed data. Biom J 62, 221-237

**Contributed Talk**

# Information sharing across genes for improved parameter estimation in concentration-response curves

**Franziska Kappenberg, Jörg Rahnenführer**
TU Dortmund University, Germany

Technologies for measuring high-dimensional gene expression values for tens of thousands of genes simultaneously are well established. In toxicology, for estimating concentration-response curves, such data can be used to understand the biological processes initiated at different concentrations. Increasing the number of concentrations or the number of replicates per concentration can improve the accuracy of the fit, but causes critical additional costs. A statistical approach to obtain higher-quality fits is to exploit similarities between high-dimensional concentration-gene expression data. This idea can also be called information sharing across genes. Parameters of the concentration-response curves can be linked, according to a priori assumptions or estimates of the distributions of the parameters, in a Bayesian framework.

Here, we consider the special case of the sigmoidal 4pLL model for estimating the curves associated with single genes, and we are interested in the EC50 value of the curve, i.e. the concentration at which the half-maximal effect is reached. This value is a parameter of the 4pLL model and can be considered a reasonable indicator for a relevant expression effect of the corresponding gene. We introduce an empirical Bayes method for information sharing across genes in this situation, by modelling the distribution of the EC50 values across all genes. Based on this distribution, for each gene a weighted mean of the individually estimated parameter and the overall mean of the estimated parameters of all genes is calculated. In other words, parameters are shrunk towards an overall mean. We evaluate our approach using several simulation studies that differ with respect to their degree of assumptions made for the distribution of the EC50 values. Finally, the method is also applied to a real gene expression dataset to demonstrate the influence of the analysis strategy on the results.

**Poster**

# A note on Rogan-Gladen estimate of the prevalence

**Barbora Kessel, Berit Lange**
Helmholtz Zentrum für Infektionsforschung, Germany

When estimating prevalence based on data obtained by an imperfect diagnostic test, an adjustment for the sensitivity and specificity of the test is desired. The test characteristics are usually determined in a validation study and are known only with an uncertainty, which should be accounted for as well. The classical Rogan-Gladen correction [4] comes with an approximate confidence interval based on normality and the delta method. However, in literature it was found to have lower than the nominal coverage when prevalence is low and both sensitivity and specificity are close to 1 [2]. In a recent simulation study [1] the empirical coverage of a nominal 95% Rogan-Gladen confidence interval was mostly below 90% and as low as 70% over a wide range of setups. These results are much worse than those reported in [2] and make Rogan-Gladen interval not recommendable in practice. Since we are interested in applying the Rogan-Gladen method to estimate seroprevalence of SARS-CoV-2 infections, like it was done e.g. in [5], we will present detailed simulation results clarifying the properties of the Rogan-Gladen method in setups with low true prevalences and high specificities as being seen in the current seroprevalence studies of SARS-CoV-2 infections, see e.g. the overview [3]. We will also take into account that in the actual studies, the final estimate is often a weighted average of prevalences in subgroups of the population. We will make recommendations when the modification of the procedure suggested by Lang and Reiczigel [2] is necessary. To conclude, we would like to note that since the uncertainties in the sensitivity and specificity estimates used for the correction influence the uncertainty of the corrected prevalence, it is highly desirable to always state not only the values of the test characteristics but also the values of their uncertainties used for the correction. Reporting also the crude uncorrected prevalences enhances the future re-use of the results e.g. in meta-analyses.

**References:**

[1] Flor M, Weiß M, Selhorst T et al (2020). BMC Public Health 20:1135, doi: 10.1186/s12889-020-09177-4

[2] Lang Z, Reiczigel J. (2014). Preventive Veterinary Medicine 113, pp. 13--22, doi: 10.1016/j.prevetmed.2013.09.015

[3] Neuhauser H, Thamm R, Buttmann-Schweiger N et al. (2020). Epid Bull 50, pp. 3--6; doi: 10.25646/7728

[4] Rogan WJ, Gladen B (1978). American Journal of Epidemiology 107(1), pp. 71--76, doi: 10.1093/oxfordjournals.aje.a112510

[5] Santos-Hövener C, Neuhauser HK, Schaffrath Rosario A et al. (2020). Euro Surveill 25(47):pii=2001752, doi: 10.2807/1560-7917.ES.2020.25.47.2001752

**Poster**

# Biometrical challenges of the Use Case of the Medical Informatics Initiative (MI-I) on „POLypharmacy, drug interActions, Risks" (POLAR_MI)

**Miriam Kesselmeier[1], Martin Boeker[2], Julia Gantner[1], Markus Löffler[3], Frank Meineke[3], Thomas Peschel[3], Jens Przybilla[3], André Scherag[1], Susann Schulze[4], Judith Schuster[3], Samira Zeynalova[3], Daniela Zöller[2]**

[1]Institute of Medical Statistics, Computer and Data Sciences, Jena University Hospital, Jena, Germany; [2]Institute of Medical Biometry and Statistics, University of Freiburg, Freiburg, Germany; [3]Institute for Medical Informatics, Statistics and Epidemiology, University Leipzig, Leipzig, Germany; [4]Geschäftsbereich Informationstechnologie, Universitätsklinikum Hamburg-Eppendorf, Hamburg, Germany

**Introduction:**

The aim of POLAR_MI is to use (and, where necessary, adapt/develop) methods and processes of the MI-I to contribute to the detection of health risks in patients with polypharmacy. Polypharmacy occurs especially in elderly patients with multi-morbidity. It is associated with an increased risk for medication errors and drug-drug or drug-disease interactions, which either reduce or intensify the desired effect of individual active substances or lead to undesired adverse drug effects. The project involves an interdisciplinary team ranging from medical informatics to pharmacy and clinical pharmacology with experts from 21 institutions, among them 13 university hospitals and their data integration centres (DICs). Here we focus on some of the biometrical challenges of POLAR_MI.

**Material and methods:**

POLAR_MI relies on the infrastructure of the DICs. The tasks of a DIC include the transfer of data from a wide range of data-providing systems, their interoperable integration and processing while ensuring data quality and data protection. Ultimately, DICs should contribute to a data sharing culture in medicine along the FAIR principles. POLAR_MI is designed to utilize (anonymous) data conforming to the MI-I core data set specification. The generic biometrical concept foresees a two-step procedure: 1) Aggregation (including analysis) of individual patient-level data locally at each DIC using distributed computing mechanisms and shared algorithms, because the security of personal data cannot be guaranteed by anonymisation alone. 2) Afterwards, combination of aggregated data across all contributing DICs. The formulation of these steps requires continuous feedback from the other working groups of POLAR_MI.

**Results:**

To practically implement the biometric concept, we have developed multiple, small iterative steps that alternate between pharma and DIC team. These steps are addressed by pilot data use projects initially limited to single DICs. The steps cover definitions of potentially drug-related events (like falls, delirium or acute renal insufficiency), active substances, outcomes and related value ranges or requirements regarding data privacy issues. Additionally, the handling of missing data, possibly non-ignorable heterogeneity between the DICs as well as inclusion and exclusion criteria for the different research questions of POLAR_MI were discussed.

**Conclusion:**

Based on the results of the pilot data use projects, analyses (approaches) of the main hypotheses of POLAR_MI will be developed. The iterative workflow and the necessary steps presented here may serve as a blue print for other projects using real world data.

**Contributed Talk**

# Analysis and sample size calculation for a conditional survival model with a binary surrogate endpoint

**Samuel Kilian, Johannes Krisam, Meinhard Kieser**
Institute of Medical Biometry and Informatics; University Heidelberg; Heidelberg, Germany

The primary endpoint in oncology is usually overall survival, where differences between therapies may only be observable after many years. To avoid withholding of a promising therapy, preliminary approval based on a surrogate endpoint is possible in certain situations (Wallach et al., 2018). The approval has to be confirmed later when overall survival can be assessed. When this is done within the same study, the correlation between surrogate endpoint and overall survival has to be taken into account for sample size calculation and analysis. This relation can be modeled by means of a conditional survival model which was proposed by Xia et al. (2014). They investigated the correlation and assessed power of the logrank test but did not develop methods for statistical testing, parameter estimation, and sample size calculation.

In this talk, a new statistical testing procedure based on the conditional model and Maximum Likelihood (ML) estimators for its parameters will be presented. An asymptotic test for survival difference will be given and an approximate sample size formula will be derived. Furthermore, an exact test for survival difference and an algorithm for exact sample size determination will be provided. Type I error rate, power, and required sample size for both newly developed tests will be determined exactly. Sample sizes will be compared to those required for the logrank test.

It will be shown that for small sample sizes the asymptotic parametric test and the logrank test exceed the nominal significance level under the conditional model. For a given sample size, the power of the asymptotic and the exact parametric test is similar, whereas the power of the logrank test is considerably lower in many situations. The other way round, the sample size needed to attain a prespecified power is comparable for the asymptotic and the exact parametric test, but considerably higher for the logrank test in many situations.

We conclude that the presented exact test performs very well under the assumptions of the conditional model and is a better choice than the asymptotic parametric test or the logrank test, respectively. Furthermore, the talk will give some insights in performing exact calculations for parametric survival time models. This provides a fast and powerful method to evaluate parametric tests for survival difference, thus facilitating the planning, conduct, and analysis of oncology trials with the option of accelerated approval.

**Contributed Talk**

# Arguments for exhuming nonnegative garrote out of grave

**Edwin Kipruto**, **Willi Sauerbrei**
Medical Center-University of Freiburg, Germany

**Background:**

The original nonnegative garrote (Breiman 1995) seems to have been forgotten despite some of its good conceptual properties. Its unpopularity is probably caused by dependence on least square estimates which does not have solution in high dimensional data and performs very poorly in high degree of multicollinearity. However, Yuan and Lin (2007) showed that nonnegative garrote is a flexible approach that can be used in combination with other estimators besides least squares such as ridge hence the aforementioned challenges can be circumvented; despite this proposal, it is hardly used in practice. Considerable attention has been given to prediction models compared to descriptive models where the aim is to summarize the data structure in a compact manner (Shmueli, 2010). Here our main interest is on descriptive modeling and as a byproduct we will present results of prediction.

**Objectives:**

To evaluate the performance of nonnegative garrote and compare results with some popular approaches using three different real datasets with low to high degree of multicollinearity and in high dimensional data

**Methods:**

We evaluated four penalized regression methods (Nonnegative garrote, lasso, adaptive lasso, relaxed lasso) and two classical variable selection methods (best subset, backward elimination) with and without post-estimation shrinkage.

**Results:**

Nonnegative garrote can be used with other initial estimators besides least squares in highly correlated data and in high dimensional datasets. Negligible differences in predictions were observed in methods while considerable differences were observed in the number of variables selected.

**Conclusion:**

To fit nonnegative garrote in highly correlated data and in high dimensional settings the proposed initial estimates can be used as an alternative to least squares estimates.

**Contributed Talk**

# Evaluation of augmentation techniques for high-dimensional gene expression data for the purpose of fitting artificial neural networks

**Magdalena Kircher, Jessica Krepel, Babak Saremi, Klaus Jung**
University of Veterinary Medicine Hannover, Foundation, Germany

**Background:**

High-throughput transcriptome expression data from DNA microarrays or RNA-seq are regularly checked for their ability to classify samples. However, with further densification of transcriptomic data and a low number of observations – due to a lack of available biosamples, prohibitive costs and ethical reasons – the ratio between the number of variables and available observations is usually very large. As a consequence, classifier performance estimated from training data often tends to be overrated and little robust. It has been demonstrated in many applications that data augmentation can improve the robustness of artificial neural networks. Data augmentation on high-dimensional gene expression data has, however, been very little studied so far.

**Methods:**

We investigate the applicability and capacity of two data augmentation approaches including generative adversarial networks (GAN), which have been widely used for augmenting image datasets. Comparison of augmentation methods is carried out in public example data sets from infection research. Besides neural networks, we evaluate the augmentation techniques on the performance of linear discriminant analysis and support vector machines.

**Results and Outlook:**

First results of a 10-fold cross validation show increased accuracy, sensitivity, specificity and predictive values when using augmented data sets compared to classifier models based on the original data only. A simple augmentation approach by mixed observations shows a similar performance as the computational more expensive approach with GANs. Further evaluations are currently running to better understand the detailed performance of the augmentation techniques.

**Contributed Talk**

# Herausforderungen der Online-Lehre und was wir gelernt haben – am Beispiel des Masterstudiengangs Medical Biometry/Biostatistics und Zertifikats Medical Data Science der Universität Heidelberg

**Marietta Kirchner, Regina Krisam, Meinhard Kieser**
Institute of Medical Biometry and Informatics, Heidelberg University, Germany

Am Institut für Medizinische Biometrie und Informatik der Universität Heidelberg wird seit 2006 der weiterbildende Masterstudiengang Medical Biometry/ Biostatistics und seit 2019 das Zertifikat Medical Data Science angeboten. Beide Programme sind berufsbegleitend, deren Lehrveranstaltungen in Blockkursen an 3 aufeinander folgenden Tagen mit mehreren 90-minütigen Einheiten stattfinden. Als im März 2020 aufgrund der COVID-19 Pandemie alle Präsenzlehrveranstaltungen der Universität Heidelberg mit sofortiger Wirkung eingestellt wurden, erforderte dies eine schnelle Umorganisation der laufenden und anstehenden Kurse, um den Studienbetrieb erfolgreich aufrecht zu erhalten. Die Universität Heidelberg stellte ein Online Curriculum bereit, welches fortlaufend angepasst wurde, sowie ein Videokonferenzsystem für synchrone Lehrveranstaltungen.

Die abrupte Unterbrechung und der schnelle Umstieg auf Online-Lehre führten zu neuen Herausforderungen, bei denen das Zurückgreifen auf bewährte Vorgehensweisen nicht gegeben war. Die praktischen Programmier-Einheiten in R und die Block-Gestaltung stellten hierbei zusätzliche Herausforderungen dar, sowohl für die Teilnehmer als auch für die Dozenten. Auch wenn das Angebot an Online Lehrveranstaltungen in den letzten Jahren stetig zugenommen hat, ist nicht umfassend untersucht, ob dies einen vergleichbaren Wert wie der traditionelle Präsenzunterricht hat und welche Voraussetzungen geschaffen werden müssen für eine erfolgreiche Lehr-/Lernsituation. Richtig umgesetzt kann Online-Lehre zu einer Leistungsverbesserung bei den Studierenden führen (Shah, 2016).

Doch was macht gute Online-Lehre aus? Gelungene Online-Lehrveranstaltungen nutzen die Vorteile der verwendeten Online-Tools aus und fördern die Kommunikation zwischen den Dozenten und Studenten (Oliver, 1999). Das zur Verfügung gestellte Videokonferenzsystem bietet verschiedene Strategien an, um eine fruchtbare Online-Lernumgebungen zu schaffen. Einführungen in die Verwendung des Videokonferenzsystem enthielten Empfehlungen zum Einsatz des Systems und zur Förderung und Gestaltung der Interaktion mit den Studierenden.

Im Vortrag wird dargestellt, welche Herausforderungen und Chancen aus Sicht der Organisatoren der Studienprogramme und der Lehrenden aufgetreten sind. Die Sicht der Lernenden wird dargestellt basierend auf durchgeführten Evaluationen und ausführlichem Feedback aus Gesprächen und E-Mails. Es werden die Erfahrungen aus zwei Semestern Online-Lehre präsentiert mit dem Fokus auf „Was haben wir gemacht, um eine erfolgreiche Vermittlung der Inhalte zu gewährleisten?“ und „Was haben wir für zukünftige Lehrveranstaltungen gelernt – Präsenz oder Online?“.

**Referenzen:**

[1] R. Oliver (1999). Exploring strategies for online teaching and learning. Distance Education, 20:240-254. DOI: 10.1080/0158791990200205

[2] D. Shah (2016). Online education: should we take it seriously? Climacteric, 19:3-6, DOI: 10.3109/13697137.2015.1115314

**Contributed Talk**

# An R package for an integrated evaluation of statistical approaches to cancer incidence projection

**Maximilian Knoll[1,2,3,4], Jennifer Furkel[1,2,3,4], Jürgen Debus[1,3,4], Amir Abdollahi[1,3,4], André Karch[5], Christian Stock[6,7]**

[1]Department of Radiation Oncology, Heidelberg University Hospital, Heidelberg, Germany; [2]Faculty of Biosciences, Heidelberg University, Heidelberg, Germany; [3]Clinical Cooperation Unit Radiation Oncology, German Cancer Research Center (DKFZ), Heidelberg, Germany; [4]German Cancer Consortium (DKTK) Core Center Heidelberg, Heidelberg, Germany; [5]Institute of Epidemiology and Social Medicine, University of Muenster, Muenster, Germany.; [6]Institute of Medical Biometry and Informatics (IMBI), University of Heidelberg, Heidelberg, Germany; [7]Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany

**Background:**

Projection of future cancer incidence is an important task in cancer epidemiology. The results are of interest also for biomedical research and public health policy. Age-Period-Cohort (APC) models, usually based on long-term cancer registry data (>20yrs), are established for such projections. In many countries (including Germany), however, nationwide long-term data are not yet available. It is unclear which statistical approach should be recommended for projections using rather short-term data.

**Methods:**

To enable a comparative analysis of the performance of statistical approaches to cancer incidence projection, we developed an R package (incAnalysis), supporting in particular Bayesian models fitted by Integrated Nested Laplace Approximations (INLA). Its use is demonstrated by an extensive empirical evaluation of operating characteristics (bias, coverage and precision) of potentially applicable models differing by complexity. Observed long-term data from three cancer registries (SEER-9, NORDCAN, Saarland) was used for benchmarking.

**Results:**

Overall, coverage was high (mostly >90%) for Bayesian APC models (BAPC), whereas less complex models showed differences in coverage dependent on projection-period. Intercept-only models yielded values below 20% for coverage. Bias increased and precision decreased for longer projection periods (>15 years) for all except intercept-only models. Precision was lowest for complex models such as BAPC models, generalized additive models with multivariate smoothers and generalized linear models with age x period interaction effects.

**Conclusion:**

The incAnalysis R package allows a straightforward comparison of cancer incidence rate projection approaches. Further detailed and targeted investigations into model performance in addition to the presented empirical results are recommended to derive guidance on appropriate statistical projection methods in a given setting.

**Contributed Talk**

# Individualizing deep dynamic models for psychological resilience data

**Göran Köber[1,2], Shakoor Pooseh[2,3], Haakon Engen[4], Andrea Chmitorz[5,6,7], Miriam Kampa[5,8,9], Anita Schick[4,10], Alexandra Sebastian[6], Oliver Tüscher[5,6], Michèle Wessa[5,11], Kenneth S.L. Yuen[4,5], Henrik Walter[12,13], Raffael Kalisch[4,5], Jens Timmer[2,3,14], Harald Binder[1,2]**

[1]Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Germany; [2]Freiburg Center of Data Analysis and Modelling (FDM), University of Freiburg, Freiburg, 79104, Germany; [3]Institute of Physics, University of Freiburg, 79104, Germany; [4]Neuroimaging Center (NIC), Focus Program Translational Neuroscience (FTN), Johannes Gutenberg University Medical Center, Mainz, 55131, Germany; [5]Leibniz Institute for Resilience Research (LIR), Mainz, 55122, Germany; [6]Department of Psychiatry and Psychotherapy, Johannes Gutenberg University Medical Center, Mainz, 55131, Germany; [7]Faculty of Social Work, Health and Nursing, University of Applied Sciences Esslingen, Esslingen, 73728, Germany; [8]Department of Clinical Psychology, University of Siegen, 57076, Germany; [9]Bender Institute of Neuroimaging (BION), Department of Psychology, Justus Liebig University, Gießen, 35394, Germany; [10]Department of Public Mental Health, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Germany; [11]Department of Clinical Psychology and Neuropsychology, Institute of Psychology, Johannes Gutenberg University, Mainz, 55131, Germany; [12]Research Division of Mind and Brain, Charité–Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Germany; [13]Berlin School of Mind and Brain, Humboldt-Universität zu Berlin, Germany; [14]CIBSS—Centre for Integrative Biological Signaling Studies, University of Freiburg, 79104, Germany

Deep learning approaches can uncover complex patterns in data. In particular, variational autoencoders (VAEs) achieve this by a non-linear mapping of data into a low-dimensional latent space. Motivated by an application to psychological resilience in the Mainz Resilience Project (MARP), which features intermittent longitudinal measurements of stressors and mental health, we propose an approach for individualized, dynamic modeling in this latent space. Specifically, we utilize ordinary differential equations (ODEs) and develop a novel technique for obtaining person-specific ODE parameters even in settings with a rather small number of individuals and observations, incomplete data, and a differing number of observations per individual. This technique allows us to subsequently investigate individual reactions to stimuli, such as the mental health impact of stressors. A potentially large number of baseline characteristics can then be linked to this individual response by regularized regression, e.g., for identifying resilience factors. Thus, our new method provides a way of connecting different kinds of complex longitudinal and baseline measures via individualized, dynamic models. The promising results obtained in the exemplary resilience application indicate that our proposal for dynamic deep learning might also be more generally useful for other application domains.

**Contributed Talk**

# The max-t Test in High-Dimensional Repeated Measures and Multivariate Designs

**Frank Konietschke**
Charite Berlin, Germany

Repeated measures (and multivariate) designs occur in a variety of different research areas. Hereby, the designs might be high-dimensional, i.e. more (possibly) dependent than independent replications of the trial are observed. In recent years, several global testing procedures (studentized quadratic forms) have been proposed for the analysis of such data. Testing global null hypotheses, however, usually does not answer the main question of practitioners, which is the specific localization of significant time points or group*time interactions. The use of max-t tests on the contrary, can provide this important information. In this talk, we discuss its applicability in such designs. In particular, we approximate the distribution of the max t-test statistic using innovative resampling strategies. Extensive simulation studies show that the test is particularly suitable for the analysis of data sets with small sample sizes . A real data set illustrates the application of the method.

**Invited Talk**

# Open questions to genetic epidemiologists

**Inke R. König**
Universität zu Lübeck, Germany

Given the rapid pace with which genomics and other - omics disciplines are evolving, it is sometimes necessary to shift down a gear to consider more general scientific questions. In this line, we can formulate a number of questions for genetic epidemiologists to ponder on. These cover the areas of reproducibility, statistical significance, chance findings, precision medicine and overlaps with related fields such as bioinformatics and data science. Importantly, similar questions are being raised in other biostatistical fields. Answering these requires to think outside the box and to learn from other, related, disciplines. From that, possible hints at responses are presented to foster the further discussion of these topics.

**Invited Talk**

# Epidemiologische Modelle in der Öffentlichkeit – mit Statistik durch die Pandemie

**Lars Koppers**
Science Media Center Germany; Department für Wissenschaftskommunikation, Karlsruher Institut für Technologie, Germany

Die Corona-Pandemie hat gezeigt, wie wichtig mathematisches und statistisches Grundwissen auch im Alltag ist. Seit Anfang 2020 werden auch in der Öffentlichkeit statistische Maßzahlen und Modelle diskutiert. Die Bandbreite reicht dabei von einfachen Meldezahlen über Mittelwerte bis zu SIR-Modellen und Simulationen von aktiven Teilchen. Aber welche Modelle und Maßzahlen helfen in welchen Situationen? Welche Schlüsse können aus einer Simulation gezogen werden und welche nicht? Und wie können komplexe Zusammenhänge so vermittelt werden, dass diese auch in der Öffentlichkeit ankommen?

Das gemeinnützige Science Media Center Germany (SMC) wurde 2015 als Intermediär zwischen Wissenschaft und Wissenschaftsjournalismus gegründet. Es stellt dazu zeitnah Einschätzungen und Zitate zu tagesaktuellen Geschehnissen aus der Wissenschaft zur Verfügung und bieten zu unübersichtlichen oder vielschichtigen Themen Expertise und Hintergrundwissen. Das SMC Lab entwickelt als Datenlabor des SMC Software und Services für die eigene Redaktion und für die journalistische Community.

Im Zuge der Corona-Pandemie wuchs der Bedarf an statistischer Expertise im Jounalismus exponentiell. Maßzahlen wie die Verdopplungszeit, der Reproduktionsfaktor R oder die Eigenschaften eines exponentiellen Wachstums müssen so erklärt werden, dass Journalist*innen dazu befähigt werden kompetent über die Pandemie zu berichten. Ein wichtiger Schwerpunkt dabei sind auch die Limitationen eines jeden Modells, schließlich mag ein exponentielles Wachstum für einen kurzen Zeitraum eine treffende Beschreibung einer Zeitreihe sein, in einer endlichen Population kommt dieses Modell aber schnell an seine Grenzen. Mit zuerst täglichen, inzwischen wöchentlichen Corona-Reports hilft das SMC die aktuelle Datenlage, wie zum Beispiel die Meldezahlen des Robert Koch-Instituts (RKI) und des DIVI-Intensivregisters einzuordnen und zu erklären. Insbesondere die Meldedaten des RKI erzeugen dabei einen hohen Erklärungsbedarf, da Meldeverzug und die Tatsache, dass es sich hier nicht um eine Zufallsstichprobe handelt, dazu verleiten, falsche Schlüsse zu ziehen.

Im Bereich der epidemiologischen Modelle wurden im vergangenen Jahr von vielen Gruppen Preprints und Paper veröffentlicht, oft begleitet von online zugänglichen Dashboards und der Pressemitteilung der zugehörigen Einrichtung. Nicht jedes neue Modell trägt allerdings zum Erkenntnisstand bei, zuweilen fehlt es an fachlicher Expertise in der Modellierung einer Pandemie, die Validierung von Prognosen ist oft unzureichend. Eine Auseinandersetzung mit der Öffentlichkeitswirkung der publizierten Arbeit ist hier notwendig, erst recht wenn dies außerhalb der üblichen Peer Review Verfahren geschieht.

**Contributed Talk**

# Predictions by random forests – confidence intervals and their coverage probabilities

**Diana Kormilez, Björn-Hergen Laabs, Inke R. König**
Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Germany

Random forests are a popular supervised learning method. Their main purpose is the robust prediction of an outcome based on a learned set of rules. To evaluate the precision of predictions their scattering and distributions are important. In order to quantify this, 95 % confidence intervals for the predictions can be generated using suitable variance estimators. However, these variance estimators may be under- or overestimated and the confidence intervals thus cover ranges either too small or too large, which can be evaluated by estimating coverage probabilities through simulations. The aim of our study was to examine coverage probabilities for two popular variance estimators for predictions made by random forests, the infinitesimal jackknife according to Wager et al. (2014) and the fixed-point based variance estimator according to Mentch and Hooker (2016). We performed a simulation study considering different scenarios with varying sample sizes and various signal-to-noise ratios. Our results show that the coverage probabilities based on the infinitesimal jackknife are lower than the desired 95 % for small data sets and small random forests. On the other hand, the variance estimator according to Mentch and Hooker (2016) leads to overestimated coverage probabilities. However, a growing number of trees yields decreasing coverage probabilities for both methods. A similar behavior was observed when using real datasets, where the composition of the data and the number of trees influence the coverage probabilities. In conclusion, we observed that the relative performance of one variance estimation method over the other depends on the hyperparameters used for training the random forest. Likewise, the coverage probabilities can be used to evaluate how well the hyper-parameters were chosen and whether the data set requires more pre-processing.

**References:**

[1] Mentch L, Hooker G (2016): Quantifying Uncertainty in Random Forests via Confidence Intervals and Hypothesis Tests. J. Mach. Learn. Res., 17:1-41.

[2] Wager S, Hastie T, Efron B (2014): Confidence Intervals for Random Forests: The Jackknife and the Infinitesimal Jackknife. J. Mach. Learn. Res., 15:1625-1651.

**Contributed Talk**

# Comparison of merging strategies for building machine learning models on multiple independent gene expression data sets

**Jessica Krepel, Magdalena Kircher, Moritz Kohls, Klaus Jung**
University of Veterinary Medicine Hannover, Foundation, Germany

**Background:**

Microarray experiments and RNA-seq experiments allow the simultaneous collection of gene expression data from several thousand genes which can be used in a wide range of biological questions. Nowadays, there are gene expression data available in public databases for many biological and medical research questions. Oftentimes, several independent studies are performed on the same or similar research question. There are several benefits of combining these studies compared to individual analyses. Several approaches for combining independent data sets of gene expression data have been proposed already in the context of differential gene expression analysis and gene set enrichment analysis. Here, we want to compare different strategies for combining independent data sets for the purpose of classification analysis.

**Methods:**

We only considered the two-group design, e.g. with class labels diseased and healthy. At different stages of the analysis, the information of the individual studies can be aggregated. We examined three different merging pipelines with regard to the stage of the analysis at which merging is conducted, namely the direct merging of the data sets (strategy A), the merging of the trained classification models (strategy B), and the merging of the classification results (strategy C). We combined the merging pipelines with different methods for classification, linear discriminant analysis (LDA), support vector machines (SVM), and artificial neural networks (ANN). Within each merging strategy, we performed a differential gene expression analysis for dimension reduction to select a set of genes that we then used as feature subset in the classification. We trained and evaluated the classification model on several data subsets in form of a 10-fold cross validation. We first performed a simulation study with pure artificial data, and secondly a study based on a real world data set from the public data repository ArrayExpress that we artificially split into two studies.

**Results:**

With respect to classification accuracy, we found that the strategy of data merging outperformed the strategy of results merging in most of our simulation scenarios with artificial data. Especially when the number of studies is high and the differentiability between the groups is low, strategy A appears as the best performing one. Strategy A performed particularly better than the other two merging approaches when four independent studies were aggregated compared to scenarios with only two independent studies.

**Contributed Talk**

# Type X Error: Is it time for a new concept?

**Cornelia Ursula Kunz**
Boehringer Ingelheim Pharma GmbH & Co. KG, Germany

A fundamental principle of how we decide between different trial designs and different test statistics is the control of error rates as defined by Neyman and Pearson, namely the type I error rate and the type II error rate. The first one controlling the probability to reject a true null hypothesis and the second one controlling the probability to not reject a false null hypothesis. When Neyman and Pearson first introduced the concepts of type I and type II error, they could not have predicted the increasing complexity of many trials conducted today and the problems that arise with them.

Modern clinical trials often try to address several clinical objectives at once, hence testing more than one hypothesis. In addition, trial designs are becoming more and more flexible, allowing to adapt ongoing trial by changing, for example, number of treatment arms, target populations, sample sizes and so on. It is also known that in some cases the adaptation of the trial leads to a change of the hypothesis being tested as for example happens when the primary endpoint of the trial is changed at an interim analysis.

While their focus was on finding the most powerful test for a given hypothesis, we nowadays often face the problem of finding the right trial design in the first place before even attempting on finding the most powerful or in some cases even just a test at all. Furthermore, when more than one hypothesis is being tested family-wise type I error control in the weak or strong sense also has to be addressed with different opinions on when we need to control for it and when we might not need to.

Based on some trial examples, we show that the more complex the clinical trial objectives, the more difficult it might be to establish a trial that is actually able to answer the research question. Often it is not sufficient or even possible to translate the trial objectives into simple hypotheses that are then being tested by some most powerful test statistic. However, when the clinical trial objectives cannot completely be addressed by a set of null hypotheses, type I and type II error might not be sufficient anymore to decide on the admissibility of a trial design or test statistic. Hence, we raise the question whether a new kind of error should be introduced.

**Invited Talk**

# Do we still need hazard ratios? (I)

**Oliver Kuß**
German Diabetes Center, Leibniz Institute for Diabetes Research at Heinrich Heine University Düsseldorf, Institute for Biometrics and Epidemiology

It is one of the phenomenons in biostatistics that regression models for continuous, binary, nominal, or ordinal outcomes almost completely rely on parametric modelling, whereas survival or time-to-event outcomes are mainly analyzed by the Proportional Hazards (PH) model of Cox, which is an essentially non-parametric method. The Cox model has become one of the most used statistical models in applied research and the original article from 1972 ranks below the top 100 papers (in terms of citation frequency) across all areas of science.

However, the Cox model and the hazard ratio have also been criticized recently. For example, researchers have been warned to use the magnitude of the HR to describe the magnitude of the relative risk, because the hazard ratio is a ratio of rates, and not one of risks. Hazard ratios, even in randomized trials, have a built-in “selection bias”, because they are conditional measures, conditioning at each time point on the set of observations which is still under risk. Finally, the hazard ratio has been criticized for being non-collapsible. That is, adjusting for a covariate that is associated with the event will in general change the HR, even if this covariate is not associated with the exposure, that is, is no confounder.

In view of these disadvantages it is surprising that parametric survival models are not preferred over the Cox model. These existed long before the Cox model, are easier to comprehend, estimate, and communicate, and, above all, do not have any of the disadvantages mentioned.

**Contributed Talk**

# Identification of representative trees in random forests based on a new tree-based distance measure

**Björn-Hergen Laabs, Inke R. König**
Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Germany

In life sciences random forests are often used to train predictive models, but it is rather complex to gain any explanatory insight into the mechanics leading to a specific outcome, which impedes the implementation of random forests in clinical practice. Typically, variable importance measures are used, but they can neither explain how a variable influences the outcome nor find interactions between variables; furthermore, they ignore the tree structure in the forest in total. A different approach is to select a single or a set of a few trees from the ensemble which best represent the forest. It is hoped that by simplifying a complex ensemble of decision trees to a set of a few representative trees, it is possible to observe common tree structures, the importance of specific features and variable interactions. Thus, representative trees could also help to understand interactions between genetic variants.

The intuitive definition of representative trees are those with the minimal distance to all other trees, which requires a proper definition of the distance between two trees. The currently proposed tree-based distance metrics[1] compare trees regarding either the prediction, the clustering in the terminal nodes, or the variables that were used for splitting. Therefore they either need an additional data set for calculating the distances or capture only few aspects of the tree architecture. Thus, we developed a new tree-based distance measure, which does not use an additional data set and incorporates more of the tree structure, by evaluating not only whether a certain variable was used for splitting in the tree, but also where in the tree it was used. We compared our new method with the existing metrics in an extensive simulation study and show that our new distance metric is superior in depicting the differences in tree structures. Furthermore, we found that the most representative tree selected by our method has the best prediction performance on independent validation data compared to the trees selected by other metrics.

**References:**

[1] Banerjee et al. (2012), Identifying representative trees from ensembles, Statistics in Medicine 31(15), 1601-16

**Contributed Talk**

# Weightloss as Safety Indicator in Rodents

**Tina Lang, Issam Ben Khedhiri**
Bayer AG, Germany

In preclinical research, the assessment of animal well-being is crucial to ensure ethical standards and compliance with guidelines. It is a tough task to define rules within which the well-being is deemed ok, and when to claim that the suffering of the animal exceeds a tolerable burden and thus, the animal needs to be sacrificed. Indicators are, e.g., food refusal, listlessness and, most outstanding, body weight.

For rodents, a popular rule states that animals that experiences > 20% body weight loss exceeds limits of tolerable suffering and has to be taken out of the experiment. However, research experiments are of highly various nature (Talbot et al., 2020). An absolute rule for all of them can lead to unnecessary deaths of lab animals that are still within reasonable limits of well-being, but for various reasons fall below the body weight limit.

An additional challenge are studies on juvenile rodents which are still within their growth phase. Here, a weight loss might not be observable, but a reduced weight gain could indicate complications. As a solution, their weight gain is routinely compared to the mean weight gain of a control group of animals. If the weight gain differs by a certain percentage, the animals are excluded from the experiment. In case of frequent weighing and small weight gains in the control group, this leads to mathematically driven exclusion of animals which are fit and healthy.

We propose a different approach of safety monitoring which firstly unify assessment for juvenile and adult animals and secondly compensate for different conditions within different experiments.

If a reasonable control group can be kept within the study design, the body weight within the control group is assumed to be lognormally distributed. Within the interval of mean log body weight plus/minus three standard deviations, about 99.73% of all control animals are expected to be found. We conclude that this interval contains acceptable body weights. As the theoretical means and standard deviations of log body weight are unknown, we checked how their empirical equivalent counterparts perform.

We investigated if the rule leaves all healthy animals in the study and only excludes suffering animals. Our data shows that it outperforms the traditional rules by far. Many animals that would have been excluded by the traditional rules can now stay in the study. Thus, the new rule supports animal welfare, and also increases the power of the experiment.

**Invited Talk**

# Truncation by death and the survival-incorporated median: What are we measuring? And why?

**Judith J. Lok[1], Qingyan Xiang[2], Ronald J. Bosch[3]**

[1]Department of Mathematics and Statistics, Boston University, United States of America; [2]Department of Biostatistics, Boston University, United States of America; [3]Center for Biostatistics in AIDS Research, Harvard University, United States of America

One could argue that if a person dies, their subsequent health outcomes are missing. On the other hand, one could argue that if a person dies, their health outcomes are completely obvious. This talk considers the second point of view, and advocates to not always see death as a mechanism through which health outcomes are missing, but rather as part of the outcome measure. This is especially useful when some people's lives may be saved by a treatment we wish to study. We will show that both the median health score in those alive and the median health score in the always-survivors can lead one to believe that there is a trade-off between survival and good health scores, even in cases where in clinical practice both the probability of survival and the probability of a good health score are better for one treatment arm. To overcome this issue, we propose the survival-incorporated median as an alternative summary measure of health outcomes in the presence of death. It is the outcome value such that 50% of the population is alive with an outcome above that value. The survival-incorporated median can be interpreted as what happens to the "average" person. The survival-incorporated median is particularly relevant in settings with non-negligible mortality. We will illustrate our approach by estimating the effect of statins on neurocognitive function.

**Invited Talk**

# How to enhance gameful learning in the STEM subjects

**Amir Madany Mamlouk**
Institute for Neuro- and Biocomputing, University of Lübeck, Germany

Playing games is fun, learning should actually be just as much fun. But at universities - in the STEM subjects in particular - it's usually not fun at all. On the contrary, many students drop out of their studies because they are not up to the requirements and cannot close existing gaps in knowledge. Others get sick in their studies because they are not up to the demands. In this lecture, I would like to raise awareness of the fact that our current study system often runs counter to all the principles of a successful game design. Furthermore, I would like to tell you in this talk about my own efforts to correct this systemic misalignment between learning at universities and gameful learning. Over the last few years, we have developed a multiple award-winning experience points-based assessment system (XPerts - From Zero to Hero) and systematically evaluated it in practice using a lecture on bioinformatics. I will illustrate this with a few examples and offer suggestions on how you can already achieve a fundamental change in the teaching and learning culture in your own courses, even with the smallest of changes.

**Contributed Talk**

# Testing Instrument Validity in Multivariable Mendelian Randomisation

**Maximilian Michael Mandl[1], Anne-Laure Boulesteix[1], Stephen Burgess[2], Verena Zuber[3]**
[1]Ludwig-Maximilians-Universität München; [2]University of Cambridge; [3]Imperial College London

Identification of causal effects in biomedical sciences is a challenging task. Most causal inference methods rely on specific assumptions which in practice may be unrealistic and too restrictive. However, Mendelian Randomisation (MR) is an instrumental variable approach that makes use of genetic variants to infer a causal effect of a risk factor on an outcome. Due to the randomisation of the genetic variants during meiosis, these are predestined instrumental variables that have the potential to naturally meet the restrictive methodological requirements. Thus, causal effects can be consistently inferred even if unobserved confounders are present. Obviously, this setting still requires the genetic variants to be independent of the outcome conditional on the risk factor and unobserved confounders, which is known as the exclusion-restriction assumption (ERA). Violations of this assumption, i.e. the effect of the instrumental variables on the outcome through a different path than the risk factor included in the model, can be caused by pleiotropy, which is a common phenomenon in human genetics. As an extension to the standard MR approach, multivariable MR includes multiple potential risk factors in one joint model accounting for measured pleiotropy. Genetic variants which deviate from the ERA appear as outliers to the MR model fit and can be detected by general heterogeneity statistics proposed in the literature. In MR analysis these are often inflated due to heterogeneity of how genetic variants exert their downstream effect on the exposures of interest, which impedes detection of outlying instruments using the traditional methods.

Removing valid instruments or keeping invalid instruments in the MR model may lead to a bias of the causal effect estimates and false positive findings. As different heterogeneity measures lead to a variety of conclusions with regard to outlying instruments, researchers face a typical decision problem, also known as researcher degrees of freedom. These free choices in the selection of valid instruments can lead to serious problems like fishing for significance.

Firstly, we demonstrate the impact of outliers and how arbitrary choices in the selection of instrumental variables can induce false positive findings in realistic simulation studies and in the analysis of real data investigating the effect of blood lipids on coronary heart disease and Alzheimer's disease. Secondly, we propose a method that corrects for overdispersion of the heterogeneity statistics in MR analysis by making use of the estimated inflation factor to correctly remove outlying instruments and therefore accounting for pleiotropic effects.

**Contributed Talk**

# Simultanes regionales Monitorieren von SARS-CoV-2 Infektionen und COVID-19 Sterblichkeit in Bayern durch die standardisierte Infektions-mortalitätsrate (sIFR)

**Kirsi Manz, Ulrich Mansmann**
Ludwig-Maximilians-Universität München, Deutschland

**Hintergrund**

Regionale Karten erlauben einen schnellen Überblick über die räumliche Verteilung des SARS-CoV-2 Infektionsgeschehens und erlauben regionale Unterschiede zu identifizieren. Zur Vermeidung falsch-positiver Signale werden Gesundheitskarten geglättet. Dies macht eine sachgerechte Interpretation der geographischen Informationen möglich.

**Ziel des Beitrags**

Wir stellen die standardisierte Infektionsmortalitätsrate (sIFR) als Maßzahl vor, mit der sich simultan das Divergieren von standardisierten COVID-19 spezifischen Infektions- und Sterberaten regional monitorieren lässt. Regionale Abweichungen beider Prozesse von einem globalen Standard erlauben eine Priorisierung regionaler Maßnahmen zwischen Infektionsschutz und Patientenversorgung.

**Materialien und Methoden**

Die regionale sIFR ist der Quotient zwischen standardisierter Mortalitäts- und Infektionsrate. Sie beschreibt um wieviel mehr die regionale Abweichung im Sterbeprozess sich von der regionalen Abweichung im Infektionsprozess unterscheidet. Die sIFR-Werte werden mittels eines bayesianischen Konvolutionsmodells geschätzt und in Karten dargestellt. Unsere Analysen verwenden die Meldedaten zum SARS-CoV-2 Geschehen in Bayern im Jahr 2020 und betrachten 4 Zeitperioden zu je drei Monaten.

**Ergebnisse und Diskussion**

Die empirische Infektionssterblichkeit in Bayern zeigt einen abfallenden Trend über die Zeitperioden. Regionen mit höheren Abweichungen im Sterben vom bayerischen Standard verglichen zum Infektionsgeschehen (sIFR > 2) sind in den ersten drei Monaten nur in der Oberpfalz zu beobachten. Im Sommer befinden sie sich im gesamten Osten, im Spätsommer/Herbst dann im Norden Bayerns. Wir zeigen regionale Veränderungen der sIFR-Werte für Bayerns Regionen über die Zeit. Damit werden Regionen identifiziert, die zusätzlich zum Management der Infektionsausbreitung Maßnahmen zur Kontrolle der Sterblichkeit benötigen.

**Poster**

# Statistical Cure of Cancer in Schleswig-Holstein

**Johann Mattutat**[1], **Nora Eisemann**[2], **Alexander Katalinic**[1,2]

[1]Institute for Cancer Epidemiology, University of Lübeck, Germany; [2]Institute of Social Medicine and Epidemiology, University of Lübeck, Germany

Cancer patients who have survived their treatment and who have been released into remission still live with the uncertainty of late recurrences of their disease. Yet, studies have shown that the observed mortality in the patient group converges against the overall population mortality after some time for most cancer entities. The amount of excess mortality and its time course can be estimated. The time point at which it falls below a defined threshold can be interpreted as “statistical cancer cure”.

This contribution shall focus on the workflow estimating the time point of statistical cancer cure. We briefly explain each step, describe design choices, and report some exemplary results for colorectal cancer. The calculations are based on data provided by the cancer registry of Schleswig-Holstein. First, a threshold for “statistical cancer cure” is defined. Then, missing information on tumor stage at diagnosis is imputed using multiple imputation. The net survival depending on cancer entity, sex, tumor stage at diagnosis (UICC) and age is estimated using a flexible excess hazard regression model implemented in the R package “mexhaz”. The excess mortality is derived using Gauss-Legendre quadrature. Finally, survival rates are estimated conditionally on the survival time since diagnosis and the defined thresholds are applied to get the estimated time point of cure.

We focus on the probability of belonging to the group that will suffer future excess mortality and define the time point at which this probability falls below 5% as time of “statistical cancer cure”. For the example of colorectal cancer (C18-C21) diagnosed in a local stage (UICC II), this probability amounts to approximately 16% at the time of diagnosis and statistical cure is reached after 4.2 years.

Results like the ones described above may support cancer patients by removing uncertainty regarding their future prognosis. As a subsequent step, comprehensive data covering most of the common cancer entities are to be generated.

**Invited Talk**

# Statistical analysis of high-dimensional biomedical data: issues and challenges in translation to medically useful results

**Lisa Meier McShane**
Division of Cancer Treatment and Diagnosis, U.S. National Cancer Institute, National Institutes of Health, USA

Successful translation of research involving high-dimensional biomedical data to medically useful results requires a research team with expertise including clinical and laboratory science, bioinformatics, computational science, and statistics. A proliferation of pubic databases and powerful data analysis tools have led to many biomedical publications reporting results suggested to have potential clinical application. However, many of these results cannot be reproduced in subsequent studies, or the findings, although meeting statistical significance criteria or other numerical performance criteria, have no clear clinical utility. Many factors have been suggested as contributors to irreproducible or clinically non-translatable biomedical research, including poor study design, analytic instability of measurement methods, sloppy data handling, inappropriate and misleading statistical analysis methods, improper reporting or interpretation of results, and on rare occasions, outright scientific misconduct. Although these challenges can arise in a variety of medical research studies, this talk will focus on research involving use of novel measurement technologies such as “omics assays” which generate large volumes of data requiring specialized expertise and computational approaches for proper management, analysis and interpretation [http://iom.edu/Reports/2012/Evolution-of-Translational-Omics.aspx]. Research team members share responsibility for ensuring that research is performed with integrity and best practices are followed to ensure reproducible results. Further, strong engagement of statisticians and other computational scientists with experts in the relevant medical specialties is critical to generation of medically interpretable and useful findings. Through a series of case studies, the many dimensions of reproducible and medically translatable omics research are explored and recommendations aiming to increase the translational value of the research output are discussed.

**Contributed Talk**

# Optimal futility stops in two-stage group-sequential gold-standard designs

**Jan Meis, Maximilian Pilz, Meinhard Kieser**
Institute of Medical Biometry and Informatics, University of Heidelberg, Heidelberg, Germany

A common critique of non-inferiority trials comparing an experimental treatment to an active control is that they may lack assay sensitivity. This denotes the ability of a trial to distinguish an effective treatment from an ineffective one. The 'gold-standard' non-inferiority trial design circumvents this concern by comparing three groups in a hierarchical testing procedure. First, the experimental treatment is compared to a placebo group in an effort to show superiority. Only if this succeeds, the experimental treatment is tested for non-inferiority against an active control group. Ethical and practical considerations require sample sizes of clinical trials to be as large as necessary, but as small as possible. These considerations come especially pressing in the gold-standard design, as patients are exposed to placebo doses while the control treatment is already known to be effective.

Group sequential trial designs are known to reduce the expected sample size under the alternative hypothesis. In their pioneer work, Schlömer and Brannath (2013) show that the gold-standard design is no exception to this rule. In their paper, they calculate approximately optimal rejection boundaries for the gold-standard design given sample size allocation ratios of the optimal single stage design. We extend their work by relaxing the constraints put on the group allocation ratios and allowing for futility stops at interim. The futility boundaries and the sample size allocation ratios will be considered as optimization parameters, together with the efficacy boundaries. This allows the investigation of the efficiency gain by including the option to stop for futility. Allowing discontinuation of a trial when faced with underwhelming results at an interim analysis has very practical implications in saving resources and sparing patients from being exposed to ineffective treatment. In the gold-standard design, these considerations are especially pronounced. There is a large incentive to prevent further patients from being exposed to placebo treatment when interim results suggest that a confirmatory result in the final analysis becomes unlikely.

Besides the extended design options, we analyse different choices of optimality criteria. The above considerations suggest that the null hypothesis also plays an important role in the judgement of the gold-standard design. Therefore, optimality criteria that incorporate the design performance under the alternative and the null hypothesis are introduced. The results of our numerical optimization procedure for this extended design will be discussed and compared to the findings of Schlömer and Brannath.

**Contributed Talk**

# Variable Importance in Random Forests in the Presence of Confounding

**Robert Miltenberger, Christoph Wies, Gunter Grieser, Antje Jahn**
University of applied sciences Darmstadt, Deutschland

Patients with a need for kidney transplantation suffer from a lack of available organ donors. Still, patients commonly reject an allocated kidney when they consider its quality to be insufficient [1]. Rejection is of major concern as it can reduce the organs quality due to prolonged ischemic time and thus its use for further patients. To better understand the association between organ quality and patient prognosis after transplantation, random survival forests will be applied to data on more than 50.000 kidney transplantations of the US organ transplantation registry. However, the US allocation process is allocating kidneys of high quality to patients with good prognosis. Thus confounding is of major concern and needs to be adressed.

In this talk, we investigate methods to address confounding in random forest analysis by using residuals from a generalized propensity score analysis. We show, that by considering the residuals instead of original variables the permutation variable importance measures refer to semipartial correlations between outcome and variable instead of correlations that are disturbed by confounder effects. This facilitates the interpretation of the variable importance measure. As our findings rely on linear models, we further investigate the approach for non-linear and non-additive models by the use of simulations.

The proposed method is used to analyse the impact of kidney quality on failure-free survival after transplantation based on the US registry data. Results are compared to other methods, that have been proposed for a better understanding and explainability of random forest analyses [2].

**References:**

[1] Husain SA et.al.: Association Between Declined Offers of Deceased Donor Kidney Allograft and Outcomes in Kidney Transplant Candidates. JAMA Netw Open. 2019; doi:10.1001/jamanetworkopen.2019.10312

[2] Paluszynska A, Przemyslaw Biecek P and Jiang Y (2020). randomForestExplainer: Explaining and Visualizing Random Forests in Terms of Variable Importance. R package version 0.10.1. https://CRAN.R-project.org/package=randomForestExplainer

**Invited Talk**

# Doubly robust estimation of adaptive dosing rules

**Erica EM Moodie**[1], **Juliana Schulz**[2]
[1]McGill University, Canada; [2]HEC Montreal, Canada

Dynamic weighted ordinary least squares (dWOLS) was proposed as a simple analytic tool for estimating optimal adaptive treatment strategies. The approach aimed to combine the double robustness of G-estimation with the ease of implementation of Q-learning, however early methodology was limited to only the continuous outcome/binary treatment setting. In this talk, I will introduce generalized dWOLS, an extension that allowed for continuous-valued treatments to estimate optimal dosing strategies, and demonstrate the approach in estimating an optimal Warfarin dosing rule.

**Invited Talk**

# Towards stronger simulation studies in statistical research

**Tim Morris**
MRC Clinical Trials Unit at University College London, UK

Simulation studies are a tool for understanding and evaluating statistical methods. They are sometimes necessary to generate evidence about which methods are suitable and – importantly – unsuitable for use, and when. In medical research, statisticians have been pivotal to the introduction of reporting guidelines such as CONSORT. The idea is that these give readers enough understanding of how a study was conducted that they could replicate the study themselves. Simulation studies are relatively easy to replicate but, as a profession, we tend to forget our fondness for clear reporting. In this talk, I will describe some common failings and make suggestions about the structure and details that help to clarify published reports of simulation studies.

**Contributed Talk**

# Interactive review of safety data during a data monitoring committee using R-Shiny

**Tobias Mütze[1], Bo Wang[2], Douglas Robinson[2]**

[1]Statistical Methodology, Novartis Pharma AG, Switzerland; [2]Scientific Computing and Consulting, Novartis Pharma AG, Switzerland

In clinical trials it is common that the safety of patients is monitored by a data monitoring committee (DMC) that operates independently of the clinical trial teams. After each review of the accumulating trial data, it is within the DMC's responsibility to decide on whether to continue or stop the trial. The data are generally presented to DMCs in a static report through tables, listing, and sometimes figures. In this presentation, we share our experiences with supplementing the safety data review with an interactive R-Shiny app. We will first present the layout and content of the app. Then, we outline the advantages of reviewing (safety) data by means of an interactive app compared to the standard review of a DMC report, namely, extensive use of graphical illustrations in addition to tables, ability to quickly change the level of detail, and to switch between study-level data and subject-level data. We argue that this leads to a robust collaborative discussion and a more complete understanding of the data. Finally, we discuss the qualification process itself of an R Shiny app and outline how the learnings may be applied to enhance standard DMC reports

**References:**

[1] Wang, W., Revis, R., Nilsson, M. and Crowe, B., 2020. Clinical Trial Drug Safety Assessment with Interactive Visual Analytics. Statistics in Biopharmaceutical Research, pp.1-12.

[2] Fleming, T.R., Ellenberg, S.S. and DeMets, D.L., 2018. Data monitoring committees: current issues. Clinical Trials, 15(4), pp.321-328.

[3] Mütze, T. and Friede, T., 2020. Data monitoring committees for clinical trials evaluating treatments of COVID-19. Contemporary Clinical Trials, 98, 106154.

[4] Buhr, K.A., Downs, M., Rhorer, J., Bechhofer, R. and Wittes, J., 2018. Reports to independent data monitoring committees: an appeal for clarity, completeness, and comprehensibility. Therapeutic innovation & regulatory science, 52(4), pp.459-468.

**Contributed Talk**

# Multi-state modeling and causal censoring of treatment discontinuations in randomized clinical trials

**Alexandra Nießl[1], Jan Beyersmann[1], Anja Loos[2]**
[1]University of Ulm, Germany; [2]Global Biostatistics, Merck KGaA, Darmstadt, Germany

The current COVID-19 pandemic and subsequent restrictions have various consequences on planned and ongoing clinical trials. Its effects on the conduct of a clinical trial create several challenges in analyzing and interpreting study data. In particular, a substantial amount of COVID-19-related treatment interruptions will affect the ability of the study to show the primary objective of the trial.

Recently, we investigated the impact of treatment discontinuations due to a clinical hold on the treatment effect of a clinical trial. A clinical hold order by the Food and Drug Administration (FDA) to the sponsor of a clinical trial is a measure to delay a proposed or to suspend an ongoing clinical investigation. The phase III clinical trial START with primary endpoint overall survival served as the motivating data example to explore implications and potential statistical approaches for a trial continuing after a clinical hold is lifted. We proposed a multistate model incorporating the clinical hold as well as disease progression as intermediate events to investigate the impact of the clinical hold on the treatment effect. The multistate modeling approach offers several advantages: Firstly, it naturally models the dependence between PFS and OS. Secondly, it could easily be extended to additionally account for time-dependent exposures. Thirdly, it provides the framework for a simple causal analysis of treatment effects using censoring. Here, we censor patients at the beginning of the clinical hold. Using a realistic simulation study informed by the START data, we showed that our censoring approach is flexible and it provides reasonable estimates of the treatment effect, which would be observed if no clinical hold has occurred. We pointed out that the censoring approach coincides with the causal g-computation formula and has a causal interpretation regarding the intention of the initial treatment.

Within the talk, we will present our multistate model approach and show our results with a focus on the censoring approach and the link to causal inference. Furthermore, we also propose a causal filtering approach. We will discuss the assumptions that have to be fulfilled for the ‘causal’ censoring or filtering to address treatment interruptions in general settings with an external time-dependent covariate inducing a time-varying treatment and, particularly, in the context of COVID-19.

**References:**

[1] Nießl, Alexandra, Jan Beyersmann, and Anja Loos. “Multistate modeling of clinical hold in randomized clinical trials.” Pharmaceutical Statistics 19.3 (2020): 262-275

**Contributed Talk**

# Over-optimism in benchmark studies and the multiplicity of analysis strategies when interpreting their results

**Christina Nießl[1], Moritz Herrmann[2], Chiara Wiedemann[1], Giuseppe Casalicchio[2], Anne-Laure Boulesteix[1]**

[1]Institute for Medical Informatics, Biometry and Epidemiology, University of Munich (Germany); [2]Department of Statistics, University of Munich (Germany)

In recent years, the need for neutral benchmark studies that focus on the comparison of statistical methods has been increasingly recognized. At the interface between biostatistics and bioinformatics, benchmark studies are especially important in research fields involving omics data, where hundreds of articles present their newly introduced method as superior to other methods.

While general advice on the design of neutral benchmark studies can be found in recent literature, there is always a certain amount of flexibility that researchers have to deal with. This includes the choice of datasets and performance metrics, the handling of missing values in case of algorithm failure (e.g., due to non-convergence) and the way the performance values are aggregated over the considered datasets. Consequently, different choices in the benchmark design may lead to different results of the benchmark study.

In the best-case scenario, researchers make well-considered design choices prior to conducting a benchmark study and are aware of this issue. However, they may still be concerned about how their choices affect the results. In the worst-case scenario, researchers could (most often subconsciously) use this flexibility and modify the benchmark design until it yields a result they deem satisfactory (for example, the superiority of a certain method). In this way, a benchmark study that is intended to be neutral may become biased.

In this paper, we address this issue in the context of benchmark studies based on real datasets using an example benchmark study, which compares the performance of survival prediction methods on high-dimensional multi-omics datasets. Our aim is twofold. As a first exploratory step, we examine how variable the results of a benchmark study are by trying all possible combinations of choices and comparing the resulting method rankings. In the second step, we propose a general framework based on multidimensional unfolding that allows researchers to assess the impact of each choice and identify critical choices that substantially influence the resulting method ranking. In our example benchmark study, the most critical choices were the considered datasets and the performance metric. However, in some settings, we observed that the handling of missing values and the aggregation of performances over the datasets can also have a great influence on the results.

**Contributed Talk**

# Evaluating the quality of synthetic SNP data from deep generative models under sample size constraints

**Jens Nußberger, Frederic Boesel, Stefan Lenz, Harald Binder, Moritz Hess**
Universitätsklinikum Freiburg, Germany

Synthetic data such as generated by deep generative models are increasingly considered for exchanging biomedical data, such as single nucleotide polymorphism (SNP) data, under privacy constraints. This requires that the employed model did sufficiently well learn the joint distribution of the data. A major limiting factor here is the number of available empirical observations which can be used for training. Until now, there is little evidence of how well the predominant generative approaches, namely variational autoencoders (VAEs), deep Boltzmann machines (DBMs) and generative adversarial networks (GANs) learn the joint distribution of the objective data under sample size constraints. Using simulated SNP data and data from the 1000 genomes project, we here provide results from an in-depth evaluation of VAEs, DBMs and GANs. Specifically, we investigate, how well pair-wise co-occurrences of variables in the investigated SNP data, quantified as odds ratios (ORs), are recovered in the synthetic data generated by the approaches. For simulated as well as the 1000 genomes SNP data, we observe that DBMs generally can recover structure for up to 300 SNPs. However, we also observe a tendency of over-estimating ORs when the DBMs are not carefully tuned. VAEs generally get the direction and relative strength of pairwise ORs right but generally under-estimate their magnitude. GANs perform well only when larger sample sizes are employed and when there are strong pairwise associations in the data. In conclusion, DBMs are well suited for generating synthetic observations for binary omics data, such as SNP data, under sample size constraints. VAEs perform superior at smaller sample sizes but are limited with respect to learning the absolute magnitude of pairwise associations between variables. GANs require large amounts of training data and likely require a careful selection of hyperparameters.

**Contributed Talk**

# Uncertainty in treatment hierarchy in network meta-analysis: making ranking relevant

**Theodoros Papakonstantinou[1,2], Georgia Salanti[1], Dimitris Mavridis[3,4], Gerta Rücker[2], Guido Schwarzer[2], Adriani Nikolakopoulou[1,2]**

[1]Institute of Social and Preventive Medicine, University of Bern, Switzerland; [2]Institute of Medical Biometry and Statistics, University of Freiburg, Germany; [3]Department of Primary Education, University of Ioannina, Ioannina, Greece; [4]Faculty of Medicine, Paris Descartes University, Paris, France

Network meta-analysis estimates all relative effects between competing treatments and can produce a treatment hierarchy from the least to the most desirable option. While about half of the published network meta-analyses report a ranking metric for the primary outcome, methodologists debate several issues underpinning the derivation of a treatment hierarchy. Criticisms include that ranking metrics are not accompanied by a measure of uncertainty or do not answer a clinically relevant question.

We will present a series of research questions related to network meta-analysis. For each of them, we will derive hierarchies that satisfy the set of constraints that constitute the research question and define the uncertainty of these hierarchies. We have developed an R package to calculate the treatment hierarchies.

Assuming a network of T treatments, we start by deriving the most probable hierarchies along with their probabilities. We derive the probabilities of each possible treatment hierarchy (T! permutations in total) by sampling from a multivariate normal distribution with relative treatment effects as means and corresponding variance-covariance matrix. Having the frequencies for each treatment hierarchy to arise, we define complex clinical questions: probability that (1) a specific hierarchy occurs, (2) a given order is retained in the network (e.g. A is better than B and B is better than C), (3) a specific triplet of quadruple of interventions is the most efficacious, (4) a treatment is in at a specific hierarchy position and (5) a treatment is in a specific or higher position in the hierarchy. These criteria can also be combined so that any number of them simultaneously holds, either of them holds or exactly one of them holds. For each defined question, we derive the hierarchies that satisfy the set criteria along with their probability. The sum of probabilities of all hierarchies that fulfill the criterion gives the probability of the criterion to hold. We extend the procedure to compare relative treatment effects against a clinically important value instead of the null effect.

We exemplify the method and its implementation using a network of four treatments for chronic obstructive pulmonary disease where the outcome of interest is mortality and is measured using odds ratio. The most probable hierarchy has a probability of 28%.

The developed methods extend the decision-making arsenal of evidence-based health care with tools that support clinicians, policy makers and patients to make better decisions about the best treatments for a given condition.

**Contributed Talk**

# NetCoMi: Network Construction and Comparison for Microbiome Data in R

**Stefanie Peschel**[1], **Christian L. Müller**[2,3,4], **Erika von Mutius**[1,5,6], **Anne-Laure Boulesteix**[7], **Martin Depner**[1]

[1]Institute of Asthma and Allergy Prevention, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany; [2]Department of Statistics, Ludwig-Maximilians-Universität München, Munich, Germany; [3]Institute of Computational Biology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany; [4]Center for Computational Mathematics, Flatiron Institute, New York, USA; [5]Dr von Hauner Children's Hospital, Ludwig-Maximilians-Universität München, Munich, Germany; [6]Comprehensive Pneumology Center Munich (CPC-M), Member of the German Center for Lung Research, Munich, Germany; [7]Institute for Medical Information Processing, Biometry and Epidemiology, Ludwig-Maximilians-Universität München, Munich, Germany

**Background:**

Network analysis methods are suitable for investigating the microbial interplay within a habitat. Since microbial associations may change between conditions, e.g. between health and disease state, comparing microbial association networks between groups might be useful. For this purpose, the two networks are constructed separately, and either the resulting associations themselves or the network's properties are compared between the two groups.

Estimating associations for sequencing data is challenging due to their special characteristics - that is, sparsity with a high number of zeros, high dimensionality, and compositionality. Several association measures taking these features into account have been published during the last decade. Furthermore, several network analysis tools, methods for comparing network properties among two or more groups as well as approaches for constructing differential networks are available in the literature. However, no unifying tool for the whole process of constructing, analyzing and comparing microbial association networks between groups is available so far.

**Methods:**

We provide the R package "NetCoMi" implementing this whole workflow starting with a read count matrix originating from a sequencing process, to network construction, up to a statement whether single associations, local network characteristics, the determined clusters, or even the overall network structure differs between the groups. For each of the aforementioned steps, a selection of existing methods suitable for the application on microbiome data is included. Especially the function for network construction contains many different approaches including methods for treating zeros in the data, normalization, computing microbial associations, and sparsifying the resulting association matrix. NetCoMi can either be used for constructing, analyzing and visualizing a single network, or for comparing two networks in a graphical as well as a quantitative manner, including statistical tests.

**Results:**

We illustrate the application of our package using a real data set from GABRIELA study [1] to compare microbial associations in settled dust from children's rooms between samples from two study centers. The examples demonstrate how our proposed graphical methods uncover genera with different characteristics (e.g. a different centrality) between the groups, similarities and differences between the clusterings, as well as differences among the associations themselfes. These descriptive findings are confirmed by a quantitative output including a statement whether the results are statistically significant.

**References:**

[1] Jon Genuneit, Gisela Büchele, Marco Waser, Katalin Kovacs, Anna Debinska, AndrzejBoznanski, Christine Strunz-Lehner, Elisabeth Horak, Paul Cullinan, Dick Heederik, et al.The gabriel advanced surveys: study design, participation and evaluation of bias.Paediatric and Perinatal Epidemiology, 25(5):436–447, 2011.

**Contributed Talk**

# Model selection for component network meta-analysis in disconnected networks: a case study

**Maria Petropoulou, Guido Schwarzer, Gerta Rücker**
Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center – University of Freiburg, Germany

Standard network meta-analysis (NMA) synthesizes direct and indirect evidence of randomized controlled trials (RCTs), estimating the effects of several competing interventions. Many healthcare interventions are complex, consisting of multiple, possibly interacting, components. In such cases, more general models, the component network meta-analysis (CNMA) models, allow estimating the effects of components of interventions.

Standard network meta-analysis requires a connected network. However, sometimes a disconnected network (two or more subnetworks) can occur when synthesizing evidence from RCTs. Bridging the gap between subnetworks is a challenging issue. CNMA models allow to “reconnect” a network with multi-component interventions if there are common components in subnetworks. Forward model selection for CNMA models, which has recently been developed, starts with a sparse CNMA model and, by adding interaction terms, ends up with a rich CNMA. By model selection, the best CNMA model is chosen based on a trade-off between goodness of fit (minimizing Cochran’s Q statistic) and connectivity.

Our aim is to check whether CNMA models for disconnected networks can validly re-estimate the results of a standard NMA for a connected network (benchmark). We applied the methods to a case study comparing 27 interventions for any adverse event of postoperative nausea and vomiting. Starting with the connected network, we artificially constructed disconnected networks in a systematic way without dropping interventions, such that the network keeps its size. We ended up with nine disconnected networks differing in network geometry, the number of included studies, and pairwise comparisons. The forward strategy for selecting appropriate CNMA models was implemented and the best CNMA model was identified for each disconnected network.

We compared the results of the best CNMA model for each disconnected network to the corresponding results for the connected network with respect to bias and standard error. We found that the results of the best CNMA models from each disconnected network are comparable with the benchmark. Based on our findings, we conclude that CNMA models, which are entirely based on RCT evidence, are a promising tool to deal with disconnected networks if some treatments have common components in different subnetworks. Additional analyses are planned to be conducted to simulated data under several scenarios for the generalization of results.

**Contributed Talk**

# Performance evaluation of a new “diagnostic-efficacy-combination trial design” in the context of telemedical interventions

**Mareen Pigorsch[1], Martin Möckel[2], Jan C. Wiemer[3], Friedrich Köhler[4], Geraldine Rauch[1]**

[1]Charité – Universitätsmedizin Berlin, Institute of Biometry and clinical Epidemiology; [2]Charité – Universitätsmedizin Berlin, Division of Emergency and Acute Medicine, Cardiovascular Process Research; [3]Clinical Diagnostics, Thermo Fisher Scientific; [4]Charité – Universitätsmedizin Berlin, Centre for Cardiovascular Telemedicine, Department of Cardiology and Angiology

**Aims:**

Telemedical interventions in heart failure patients intend to avoid unfavourable, treatment-related events by an early, individualized care, which reacts to the current patients need. However, telemedical support is an expensive intervention and only patients with a high risk for unfavourable follow-up events will profit from telemedical care. Möckel et al. therefore adapted a “diagnostic-efficacy-combination design” which allows to validate a biomarker and investigate a biomarker-selected population within the same study. For this, cut-off values for the biomarkers were determined based on the observed outcomes in the control group to define a high-risk subgroup. This defines the diagnostic design step. These cut-offs were subsequently applied to the intervention and the control group to identify the high-risk subgroup. The intervention effect is then evaluated by comparison of these subgroups. This defines the efficacy design step. So far, it has not been evaluated if this double use of the control group for biomarker validation and efficacy comparison leads to a bias in treatment effect estimation. In this methodological research work, we therefore want to evaluate whether the “diagnostic-efficacy-combination design” leads to biased treatment effect estimates. If there is a bias, we further want to analyse its impact and the parameters influencing its size.

**Methods:**

We perform a systematic Monte-Carlo simulation study to investigate potential bias in various realistic trial scenarios that mimic and vary the true data of the published TIM-HF2 Trial. In particular we vary the event rates, the sample sizes and the biomarker distributions.

**Results:**

The results show, that indeed the proposed design leads to some bias in the effect estimators, indicating an overestimation of the effect. But this bias is relatively small in most scenarios. The larger the sample size, the more the event rates differ in the control and the intervention group and the better the biomarker can separate the high-risk from the low-risk patients, the smaller is the resulting relative bias.

**Conclusions:**

The “diagnostic-efficacy-combination design” can be recommended for clinical applications. We recommend ensuring a sufficient large sample size.

**Reference:**

[1] Möckel M, Koehler K, Anker SD, Vollert J, Moeller V, Koehler M, Gehrig S, Wiemer JC, von Haehling S, Koehler F. Biomarker guidance allows a more personalized allocation of patients for remote patient management in heart failure: results from the TIM-HF2 trial. Eur J Heart Fail. 2019;21(11):1445-58.

**Contributed Talk**

# Opportunities and limits of optimal group-sequential designs

**Maximilian Pilz[1], Carolin Herrmann[2,3], Geraldine Rauch[2,3], Meinhard Kieser[1]**

[1]Institute of Medical Biometry and Informatics - University of Heidelberg, Germany; [2]Institute of Biometry and Clinical Epidemiology - Charité University Medicine Berlin, Germany; [3]Berlin Institute of Health

Multi-stage designs for clinical trials are becoming increasingly popular. There are two main reasons for this development. The first is the flexibility to modify the study design during the ongoing trial. This possibility is highly beneficial to avoid the failure of trials whose planning assumptions were enormously wrong. However, an unplanned design modification mid-course can also be performed in a clinical trial that has initially been planned without adaptive elements as long as the conditional error principle is applied. The second reason for the popularity of adaptive designs is the performance improvement that arises by applying a multi-stage design. For instance, an adaptive two-stage design can enormously reduce the expected sample size of a trial compared to a single-stage design.

With regard to this performance reason, a two-stage design can entirely be pre-specified before the trial starts. While this still leaves open the option to modify the design, it is preferred by regulatory authorities. Recent work treats the topic of optimal adaptive designs. While those show the best possible performance, it may be difficult to communicate them to a practitioner and to outline them in a study protocol.

To overcome this problem, simpler optimal group-sequential designs may be an option worth to be considered. Those only consist of two sample sizes (stage one and stage two) and three critical values (early futility, early efficacy, final analysis). Thus, they can easily be described and communicated.

In this talk, we present a variety of examples to investigate whether optimal group-sequential designs are a valid approximation of optimal adaptive designs. We elaborate design properties that can be fulfilled by optimal group-sequential designs without considerable performance deterioration and describe situations where an optimal adaptive design may be more appropriate. Furthermore, we give recommendations of how to specify an optimal two-stage design in the study protocol in order to motivate their application in clinical trials.

**Contributed Talk**

# The iBikE Smart Learner: evaluation of an interactive web-based learning tool to specifically address statistical misconceptions

**Sophie K. Piper[1,2], Ralph Schilling[1,2], Oliver Schweizerhof[1,2], Anne Pohrt[1,2], Dörte Huscher[1,2], Uwe Schöneberg[1,2], Eike Middell[3], Ulrike Grittner[1,2]**

[1]Institute of Biometry and Clinical Epidemiology, Charité - Universitätsmedizin Berlin, Charitéplatz 1, D-10117 Berlin, Germany; [2]Berlin Institute of Health (BIH), Anna-Louisa-Karsch Str. 2, 10178 Berlin, Germany; [3]Dr. Eike Middell, Moosdorfstr. 4, 12435 Berlin

### Background

Statistics is often an unpopular subject for medical students and researchers. However, methodological skills are essential for the correct interpretation of research results and thus for the quality of research in general. Understanding statistical concepts in particular plays a central role. In standard medical training, relatively little attention is paid to the development of these competencies, so that researching physicians (from students to professors) often have deficits and misconceptions.

The most typical example is the incorrect interpretation of the p-value. Misconceptions lead to misinterpretations of what statistics can do and where certain methods reach their limits. Therefore, methods are misused and/or results are misinterpreted, which in turn can have consequences for further research and ultimately for patients.

### Methods

We developed a learning tool called the “iBikE-Smart Learner“ - an interactive, web-based teaching program similar to the AMBOSS learning software for medical students. It is designed to address common misconceptions in statistics in a targeted (modular) manner and provides teaching elements adapted to the individual knowledge and demand of the user.

Specifically, we were able to complete the first module “Statistical misconceptions about the p-value”. This module consists of a self-contained set of multiple-choice questions directly addressing common misconceptions about the p-value based on typical examples in medical Research. A first (beta) version of the “iBikE-Smart Learner” was already available at the end of October 2019 and has been tested internally by experienced staff members of our institute.

In November 2020, we started a randomized controlled trial among researchers at the Charité to evaluate this first module. We plan to recruit 100 participants. Primary outcome is the overall performance rate which will be compared between users randomized to the full version of the tool and those randomized to the control version that has all teaching features turned off. Additionally, self-reported statistical literacy before and after using the tool as well as a subjective evaluation of the tools’ usefulness were assessed.

### Results

Until submission of this abstract, 30 participants have been recruited for the ongoing randomized controlled evaluation study. We plan to promote the iBikE-Smart Learner and show results of the evaluation study at the Charité.

### Conclusions

We developed and evaluated a first module of the “iBikE-Smart Learner” as a web-based teaching tool addressing common misconceptions about the p-value.

**Contributed Talk**

# Statistical Review of Animal trials in Ethics Committees – A Guideline

**Sophie K. Piper**[1,2], **Dario Zocholl**[1,2], **Robert Röhle**[1,2], **Andrea Stroux**[1,2], **Ulf Tölch**[2], **Frank Konietschke**[1,2]

[1]Institute of Biometry and Clinical Epidemiology, Charité - Universitätsmedizin Berlin, Charitéplatz 1, D-10117 Berlin, Germany; [2]Berlin Institute of Health (BIH), Anna-Louisa-Karsch Str. 2, 10178 Berlin, Germany

Any experiment or trial involving living organism requires ethical review and agreements. Beyond reviewing medical need and goals of the trial, statistical planning of the design and sample size computations are key review criteria. Errors made in the statistical planning phase can have severe consequences on both the results and conclusions drawn from a trial. Moreover, wrong conclusions therof might proliferate and impact future trials—a rather unethical outcome of any research. Therefore, any trial must be efficient in both a medical and statistical way in answering the questions of interests to be considered as “ethically approvable”.

For clinical trials, ethical review boards are well established. This is, however, not the case for pre-clinical and especially animal trials. While ethical review boards are established within each local authority of animal welfare, most of them do not have an appointed statistician. Moreover, unified standards or guidelines on statistical planning and reporting thereof are currently missing for pre-clinical trials.

It is the aim of our presentation to introduce and discuss

i) the need for proper statistical reviews of animal trials,

ii) a guideline of mandatory ethical review criteria, involving blinding and randomization, and

iii) the need to distinguish the planning of exploratory studies from confirmatory studies in pre-clinical research.

Our statistical criteria for ethical reviews of animal trials have been implemented in a form sheet that has been used from the Landesamt für Gesundheit und Soziales (local authority of animal welfare) in Berlin since 2019. It is online available at https://www.berlin.de/lageso/gesundheit/veterinaerwesen/tierversuche/.

**Invited Talk**

# Statistical Issues in Confirmatory Platform Trials

**Martin Posch, Elias Meyer, Franz König**
Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Austria

Adaptive platform trials provide a framework to simultaneously study multiple treatments in a disease. They are multi-armed trials where interventions can enter and leave the platform based on interim analyses as well as external events, for example, if new treatments become available [3]. The attractiveness of platform trials compared to separate parallel group trials is not only due to operational aspects as a joint trial infrastructure and more efficient patient recruitment, but results also from the possibility to share control groups, to efficiently prune non-efficacious treatments, and to allow for direct comparisons between experimental treatment arms [2]. However, the flexibility of the framework also comes with challenges for statistical inference and interpretation of trial results such as the adaptivity of platform trials (decisions on the addition or dropping of arms cannot be fully pre-specified and may have an impact on the recruitment for the current trial arms), multiplicity issues (due to multiple interventions, endpoints, subgroups and interim analyses) and the use of shared controls (which may be non-concurrent controls or control groups where the control treatment changes over time). We will discuss current controversies and the proposed statistical methodology to address these issues [1,3,4]. Furthermore, we give an overview of the IMI project EU-PEARL (Grant Agreement no. 853966) that aims to establish a general framework for platform trials, including the necessary statistical and methodological tools.

**References:**

[1] Collignon O., Gartner C., Haidich A.-B., Hemmings R.J., Hofner B., Pétavy F., Posch M., Rantell K., Roes K., Schiel A. Current Statistical Considerations and Regulatory Perspectives on the Planning of Confirmatory Basket, Umbrella, and Platform Trials. Clinical Pharmacology & Therapeutics 107(5), 1059–1067, (2020)

[2] Collignon O., Burman C.F., Posch M., Schiel A.Collaborative platform trials to fight COVID-19: methodological and regulatory considerations for a better societal outcome. Clinical Pharmacology & Therapeutics (to appear)

[3] Meyer E.L., Mesenbrink P., Dunger-Baldauf C., Fülle H.-J., Glimm E., Li Y., Posch M., König F. The Evolution of Master Protocol Clinical Trial Designs: A Systematic Literature Review. Clinical Therapeutics 42(7), 1330–1360, (2020)

[4] Posch, M., & König, F. (2020). Are p-values Useful to Judge the Evidence Against the Null Hypotheses in Complex Clinical Trials? A Comment on “The Role of p-values in Judging the Strength of Evidence and Realistic Replication Expectations”. Statistics in Biopharmaceutical Research, 1-3, (2002)

**Contributed Talk**

# A cautionary tale on using imputation methods for inference in a matched pairs design.

**Burim Ramosaj, Lubna Amro, Markus Pauly**
TU Dortmund University, Germany

Imputation procedures in biomedical fields have turned into statistical practice, since further analyses can be conducted ignoring the former presence of missing values. In particular, non-parametric imputation schemes like the random forest or a combination with the stochastic gradient boosting have shown favorable imputation performance compared to the more traditionally used MICE procedure. However, their effect on valid statistical inference has not been analyzed so far. This gap is closed by investigating their validity for inferring mean differences in incompletely observed pairs while opposing them to a recent approach that only works with the given observations at hand. Our findings indicate that machine learning schemes for (multiply) imputing missing values heavily inflate type-I-error in small to moderate matched pairs, even after modifying the test statistics using Rubin's multiple imputation rule. In addition to an extensive simulation study, an illustrative data example from a breast cancer gene study has been considered.

**Contributed Talk**

# Ranking Procedures for the Factorial Repeated Measures Design with Missing Data - Estimation, Testing and Asymptotic Theory

**Kerstin Rubarth, Frank Konietschke**
Charité Berlin, Germany

A commonly used design in health, medical and biomedical research is the repeated measures design. Often, a parametric model is used for the analysis of such data. However, if sample size is rather small or if data is skewed or is on an ordinal scale, a nonparametric approach would fit the data better than a classic parametric approach, e.g. linear mixed models. Another issue, that naturally arises when dealing with clinical or pre-clinical data, is the occurrence of missing data. Most methods can only use a complete data set, if no imputation technique is applied. The newly developed ranking procedure is a flexible method for general non-normal, ordinal, ordered categorical and even binary data and uses in case of missing data all available information instead of only the information obtained from complete cases. The hypotheses are defined in terms of the nonparametric relative effect and can be tested by using quadratic test procedures as well as the multiple contrast test procedure. Additionally, the framework allows for the incorporation of clustered data within the repeated measurements. An example for clustered data are animal studies, where several animals share the same cage and are therefore clustered within a cage. Simulation studies indicate a good performance in terms of the type-I error rate and the power under different alternatives with a missing rate up to 30%, also under non-normal data. A real data example illustrates the application of the proposed methodology.

**Contributed Talk**

# Independent Censoring in Event-Driven Trials with Staggered Entry

**Jasmin Rühl**
Universitätsmedizin Göttingen, Germany

In the pharmaceutical field, randomised clinical trials with time-to-event endpoints are frequently stopped after a pre-specified number of events has been observed. This practice leads to dependent data and non-random censoring, though, which can generally not be solved by conditioning on the underlying baseline information.

If the observation period starts at the same time for all of the subjects, the assumption of independent censoring in the counting process sense is valid (cf. Andersen et al., 1993, p. 139), and the common methods for analysing time-to-event data can be applied. The situation is not as clear in case that staggered study entry is considered, though. We demonstrate that the study design at hand indeed entails general independent censoring in the sense of Andersen et al.

By means of simulations, we further investigate possible consequences of employing techniques such as the non-parametric bootstrap that make the more restrictive assumption of random censoring. The results indicate that the dependence in event-driven data with staggered entry is generally too weak to affect the outcomes; however, in settings where only few occurrences of the regarded event are observed, the implications become clearer.

**Contributed Talk**

# A Nonparametric Bayesian Model for Historical Control Data in Reproductive Toxicology

**Ludger Sandig[1], Bernd Baier[2], Bernd-Wolfgang Igl[3], Katja Ickstadt[4]**

[1]Fakultät Statistik, Technische Universität Dortmund; [2]Reproductive Toxicology, Nonclinical Drug Safety, Boehringer Ingelheim Pharma GmbH & Co. KG; [3]Non-Clinical Statistics, Biostatistics and Data Sciences Europe, Boehringer Ingelheim Pharma GmbH & Co. KG; [4]Lehrstuhl für mathematische Statistik und biometrische Anwendungen, Fakultät Statistik, Technische Universität Dortmund

Historical control data are of fundamental importance for the interpretation of developmental and reproductive toxicology studies. Modeling such data presents two challenges: Outcomes are observed on different measurement scales (continuous, counts, categorical) and on multiple nested levels (fetuses within a litter, litters within a group, groups within a set of experiments). We propose a nonparametric Bayesian approach to tackle both of them. By using a hierarchical Dirichlet process mixture model we can capture the dependence structure of observables both within and between litters. Additionally we can accommodate an arbitrary number of variables on arbitrary measurement scales at the fetus level, e.g. fetus weight (continuous) and malformation status (categorical). In a second step we extend the model to incorporate observables at higher levels in the hierarchy, e.g. litter size or maternal weight. Inference in these models is possible using Markov Chain Monte Carlo (MCMC) techniques which we implemented in R. We illustrate our approach on several real-world datasets.

**Invited Talk**

# On recent progress of topic groups and panels

**Willi Sauerbrei[1], Michal Abrahamowicz[2], Marianne Huebner[3], Ruth Keogh[4]**

[1]Medical Center – University of Freiburg, Germany, [2]McGill, Montreal, Canada, [3]Michigan State University, East Lansing, USA, [4]London School of Hygiene and Tropical Medicine, UK

Observational studies present researchers with a number of analytical challenges, related to both: complexity of the underlying processes and imperfections of the available data (e.g. unmeasured confounders, missing data, measurement errors). Whereas many methods have been proposed to address specific challenges, there is little consensus regarding which among the alternative methods are preferable for what types of data. Often, there is also lack of solid evidence concerning systematic validation and comparisons of the performance of the methods.

To address these complex issues, the STRATOS initiative was launched in 2013. In 2021, STRATOS involves more than 100 researchers from 19 countries worldwide with background in biostatistical and epidemiological methods. The initiative has 9 Topic Groups (TG), each focusing on a different set of 'generic' analytical challenges (e.g. measurement errors or survival analysis) and 11 panels (e.g. publications, simulation studies, visualisation) co-ordinate it, to share best research practices and to disseminate research tools and results from the work of the TGs.

We will provide a short overview of recent progress, point to some research urgently needed and emphasize the importance of knowledge translation.More details are provided in short reports from all TGs and some panels which are regular contributions in the Biometric Bulletin, the newsletter of the International Biometric Society (https://stratos-initiative.org/publications), since issue 3 from 2017.

**Contributed Talk**

# Investigating treatment-effect modification by a continuous covariate in IPD meta-analysis: an approach using fractional polynomials

**Willi Sauerbrei**[1], **Patrick Royston**[2]
[1]Medical Center - University of Freiburg, Germany; [2]MRC Clinical Trials Unit at UCL, London, UK

**Context:**

In clinical trials, there is considerable interest in investigating whether a treatment effect is similar in all patients, or that some prognostic variable indicates a differential response to treatment. To examine this, a continuous predictor is usually categorised into groups according to one or more cutpoints. Several weaknesses of categorisation are well known.

**Objectives:**

To avoid the disadvantages of cutpoints and to retain full information, it is preferable to keep continuous variables continuous in the analysis. The aim is to derive a statistical procedure to handle this situation when individual patient data (IPD) are available from several studies.

**Methods:**

For continuous variables, the multivariable fractional polynomial interaction (MFPI) method provides a treatment effect function (TEF), that is, a measure of the treatment effect on the continuous scale of the covariate (Royston and Sauerbrei, Stat Med 2004, 2509-25). MFPI is applicable to most of the popular regression models, including Cox and logistic regression. A meta-analysis approach for averaging functions across several studies has been proposed (Sauerbrei and Royston, Stat Med 2011, 3341-60). A first example combining these two techniques (called metaTEFs) was published (Kasenda et al, BMJ Open 2016; 6:e011148). Another approach called meta-stepp was proposed (Wang et al, Stat Med 2016, 3704-16). Using the data from Wang (8 RCTs in patients with breast cancer) we will illustrate various issues of our metaTEFs approach.

**Results and Conclusions:**

We used metaTEFs to investigate a potential treatment effect modifier in a meta-analysis of IPD from eight RCTs. In contrast to cutpoint-based analyses, the approach avoids several critical issues and gives more detailed insight into how the treatment effect is related to a continuous biomarker. MetaTEFs retains the full information when performing IPD meta-analyses of continuous effect modifiers in randomised trials. Early experience suggests it is a promising approach.

**Contributed Talk**

# Distributed Computation of the AUROC-GLM Confidence Intervals Using DataSHIELD

**Daniel Schalk[1], Stefan Buchka[2], Ulrich Mansmann[2], Verena Hoffmann[2]**

[1]Department of Statistics, LMU Munich; [2]The Institute for Medical Information Processing, Biometry, and Epidemiology, LMU Munich

Distributed calculation protects data privacy without ruling out complex statistical analyses. Individual data stays in local databases invisible to the analyst who only receives aggregated results. A distributed algorithm that calculates a ROC curve, its AUC estimate with confidence interval is presented to evaluate a therapeutic decision rule. It will be embedded in the DataSHIELD framework [1].

Starting point is the ROC-GLM approach by Pepe et al. [2]. The additivity of the Fisher information matrix, of the score vector, and of the CI proposed by DeLong [3] to aggregate intermediate results allows to design a distributed algorithm to calculate estimates of the ROC-GLM, its AUC, and CI.

We simulate scores and labels (responses) to create AUC values within the range of [0.5, 1]. The size of individual studies is uniformly distributed on [100, 2500] while the percentage of treatment-response covers [0.2,0.8]. Per scenario, 10000 studies are produced. Per study, the AUC is calculated within a non-distributed empiric as well as a distributed setting. The difference in AUC between both approaches is independent of the number of distributed components and is within the range of [-0.019, 0.013]. The boundaries of bootstrapped CIs in the non-distributed empirical setting are close to those in the distributed approach with the CI of DeLong: Range of differences in the lower boundary [-0.015, 0.03]; range of the upper boundary deviations [-0.012, 0.026].

The distributed algorithm allows anonymous multicentric validation of the discrimination of a classification rules. A specific application is the audit use case within the MII consortium DIFUTURE (difuture.de). The multicentric prospective ProV-AL-MS study (DRKS: 00014034) on patients with newly diagnosed relapsing-remitting multiple sclerosis provides the data for a privacy-protected validation of a treatment decision score (also developed by DIFUTURE) regarding discrimination between good and insufficient treatment response. The simulation results demonstrate that our algorithm is suitable for the planned validation. The algorithm is implemented in R to be used within DataSHIELD. It will be made publicly available.

**References:**

[1] Amadou Gaye et al (2014). DataSHIELD: taking the analysis to the data, not the data to the analysis. International Journal of Epidemiology

[2] Pepe, M. S. (2003). The statistical evaluation of medical tests for classification and prediction. Medicine.

[3] DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach.Biometrics, pages 837–845.

**Invited Talk**

# Semiparametric Sensitivity Analysis: Unmeasured Confounding in Observational Studies

**Daniel Scharfstein**
Department of Population Health Sciences, University of Utah School of Medicine

Establishing cause-effect relationships from observational data often relies on untestable assumptions. It is crucial to know whether, and to what extent, the conclusions drawn from non-experimental studies are robust to potential unmeasured confounding. In this paper, we focus on the average causal effect (ACE) as our target of inference. We build on the work of Franks et al. (2019) and Robins et al. (2000) by specifying non-identified sensitivity parameters that govern a contrast between the conditional (on measured covariates) distributions of the outcome under treatment (control) between treated and untreated individuals. We use semi-parametric theory to derive the non-parametric efficient influence function of the ACE, for fixed sensitivity parameters. We utilize this influence function to construct a one-step, split-sample bias-corrected estimator of the ACE. Our estimator depends on semi-parametric models for the distribution of the observed data; importantly, these models do not impose any restrictions on the values of sensitivity analysis parameters. We establish that our estimator has $\sqrt{n}$ asymptotics. We utilize our methodology to evaluate the causal effect of smoking during pregnancy on birth weight. We also evaluate the performance of estimation procedure in a simulation study. This is joint work with Razieh Nabi, Edward Kennedy, Ming-Yueh Huang, Matteo Bonvini and Marcela Smid.

**Poster**

# DNT: An R package for differential network testing, with an application to intensive care medicine

**Roman Schefzik, Leonie Boland, Bianka Hahn, Thomas Kirschning, Holger Lindner, Manfred Thiel, Verena Schneider-Lindner**

Medical Faculty Mannheim, Heidelberg University, Germany

Statistical network analyses have become popular in many scientific disciplines, where a specific and important task is to test for significant differences between two networks. In our R package DNT, which will be made available at https://github.com/RomanSchefzik/DNT, we implement an overall frame for differential network testing procedures that differ with respect to (1) the network estimation method (typically based on specific concepts of association) and (2) the network characteristic employed to measure the difference. Using permutation-based tests with variants for paired and unpaired settings, our approach is general and applicable to various overall, node-specific or edge-specific network difference characteristics. Moreover, tools for visual comparison of two networks are implemented. Along with the package, we provide a corresponding user-friendly R Shiny application.

Exemplarily, we demonstrate the usefulness of our package in a novel application to a specific issue in intensive care medicine. In particular, we show that statistical network comparisons based on parameters representing the main organ systems are beneficial for the evaluation of the prognosis of critically ill patients in the intensive care unit (ICU), using patient data from the surgical ICU of the University Medical Centre Mannheim, Germany. We specifically consider both cross-sectional comparisons between a non-survivor and a survivor group (identified from the electronic medical records by using a combined risk set sampling and propensity score matching) and longitudinal comparisons at two different, clinically relevant time points during the ICU stay: first, after admission, and second, at an event stage prior to death in non-survivors or a matching time point in survivors. While specific outcomes depend on the considered network estimation method and network difference characteristic, there are however some overarching observations. For instance, we overall discover relevant changes of organ system interactions in critically ill patients in the course of ICU treatment in that while the network structures at admission stage tend to look fairly similar among survivors and non-survivors, the corresponding networks at event stage differ substantially. In particular, organ system interactions appear to stabilize for survivors, while they do not or even deteriorate for non-survivors. Moreover, on an edge-specific level, a positive association between creatinine and C-reactive protein is typically present in all the considered networks except for the non-survivor networks at the event stage.

**Contributed Talk**

# Statistical analysis of Covid-19 data in Rhineland-Palatinate

**Markus Schepers**[1], **Konstantin Strauch**[1], **Klaus Jahn**[3], **Philipp Zanger**[2], **Emilio Gianicolo**[1]
[1]IMBEI Unimedizin Mainz, Germany; [2]Institut für Hygiene und Infektionsschutz Abteilung Humanmedizin, Landesuntersuchungsamt; [3]Gesundheitsministerium (MSAGD)

In this ongoing project we study the infection dynamics and settings of Covid-19 in Rhineland-Palatinate: what are the most common infection pathways? How does the virus typically spread?

Our analysis is based on data of all reported cases (positively tested individuals) in Rhineland-Palatinate during a specific time period, including at least 17 August - 10 November 2020. Around 20% of the reported cases have been traced to an infection cluster. This leads to a second data set of infection clusters, whose observation variables include size of the infection cluster and infection setting (such as `private household' or `restaurant'). In line with previous studies, we found that the majority of infection clusters occurs in `private households' (including gatherings where multiple households are involved). Therefore, we are collecting additional information for a stratified sample of infection clusters with infection setting `private household'. Here, the stratification is according to counties (Landkreise) with separate public health departments (Gesundheitsämter) and size of the infection cluster. We developed a questionnaire whose responses will provide the additional information. The questionnaire contains questions on contact persons, specific occasions and activities promoting the spread of the virus. We calculate descriptive statistics such as mean, median, standard deviation, min and max of the quantities of interest.

Results and observations so far include: Cities have a higher prevalence of Covid-19 cases than the countryside. Most of the infection clusters are local rather than over-regional. We also observe a phenomenon often called over-dispersion or super-spreading, meaning that a relatively small number of individuals and clusters is responsible for the majority of all infection transmissions.

**Contributed Talk**

# Pgainsim: A method to assess the mode of inheritance for quantitative trait loci in genome-wide association studies

**Nora Scherer[1], Peggy Sekula[1], Peter Pfaffelhuber[2], Anna Köttgen[1], Pascal Schlosser[1]**

[1]Institute of Genetic Epidemiology, Faculty of Medicine and Medical Center - University of Freiburg, Germany; [2]Faculty of Mathematics and Physics, University of Freiburg, Germany

**Background:**

When performing genome-wide association studies (GWAS) conventionally an additive genetic model is used to explore whether a SNP is associated with a quantitative trait regardless of the actual mode of inheritance (MOI). Recessive and dominant genetic models are able to improve statistical power to identify non-additive variants. Moreover, the actual MOI is of interest for experimental follow-up projects. Here, we extend the concept of the p-gain statistic [1] to decide whether one of the three models provides significantly more information than the others.

**Methods:**

We define the p-gain statistic of a genetic model by the comparison of the association p-value of the model with the smaller of the two p-values of the other models. Considering the p-gain as a random variable depending on a trait and a SNP in Hardy-Weinberg equilibrium under the null hypothesis of no genetic association we show that the distribution of the p-gain statistic depends only on the allele frequency (AF).

To determine critical values where the opposing modes can be rejected, we developed the R-package pgainsim (https://github.com/genepi-freiburg/pgainsim). First, the p-gain is simulated under the null hypothesis of no genetic association for a user-specified study size and AF. Then the critical values are derived as the observed quantiles of the empirical density of the p-gain. For applications with extensive multiple testing, the R-package provides an extension of the empirical critical values by a log-linear interpolation of the quantiles.

**Results:**

We tested our method in the German Chronic Kidney Disease study with urinary concentrations of 1,462 metabolites with the goal to identify non-additive metabolite QTLs. For each metabolite we conducted a GWAS under the three models and identified 119 independent mQTLs for which pval_rec or pval_dom<4.6e-11 and pval_add>min(pval_rec,pval_dom). For 38 of these, the additive modelling was rejected based on the p-gain statistics after a Bonferroni adjustment for 1 Mio*549*2 tests. These included the LCT locus with a known dominant MOI, as well as several novel associations. A simulation study for additive and recessive associations with varying effect sizes evaluating false positive and false negative rates of the approach is ongoing.

**Conclusion:**

This new extension of the p-gain statistic allows for differentiating MOIs for QTLs considering their AF and the study sample size, even in a setting with extensive multiple testing.

**References:**

[1] Petersen, A. et al. (2012) On the hypothesis-free testing of metabolite ratios in genome-wide and metabolome-wide association studies. BMC Bioinformatics 13, 120.

**Contributed Talk**

# Precision medicine in action – the FIRE3 NGS study

**Laura Schlieker[1], Nicole Krämer[1], Volker Heinemann[2,3], Arndt Stahler[4], Sebastian Stintzing[3,4]**

[1]Staburo GmbH, Germany; [2]Department of Medicine III, University Hospital, University of Munich, Germany; [3]DKTK, German Cancer Consortium, German Cancer Research Centre (DKFZ); [4]Medical Department, Division of Hematology, Oncology and Tumor Immunology (CCM), Charité Universitaetsmedizin Berlin

The choice of the right treatment for patients based on their individual genetic profile is of utmost importance in precision medicine. To identify potential signals within the large number of biomarkers it is mandatory to define criteria for signal detection beforehand and apply appropriate statistical models in the setting of high dimensional data.

For the identification of predictive and prognostic genetic variants as well as tumor mutational burden (TMB) in patients with metastatic colorectal cancer, we derived the following pre-defined and hierarchical criteria for signal detection.

a) All biomarkers identified via a multivariate variable selection procedure

b) If a) reveals no signal, all biomarkers with adjusted p-value ≤ 0.157

c) If neither a) nor b) reveals signals, the top 5 biomarkers according to sorted, adjusted p-value

Regularized regression models were used for variable selection, and the stability of the selection process was quantified and visualized. Selected biomarkers were analyzed in terms of their predictive potential on a continuous scale.

With our analyses we confirmed the predictive potential of several already known biomarkers and identified additional promising candidate variants. Furthermore, we identified TMB as a potential prognostic biomarker with a trend towards prolonged survival for patients with high TMB.

Our analyses were supported by power simulations for the variable selection method, assuming different prevalences of biomarkers, numbers of truly predictive biomarkers and effect sizes.

**Gustav-Adolf-Lienert Laureate**

# Netboost: Network Analysis Improves High-Dimensional Omics Analysis Through Local Dimensionality Reduction

**Pascal Schlosser[1,2], Jochen Knaus[2], Maximilian Schmutz[3], Konstanze Döhner[4], Christoph Plass[5], Lars Bullinger[6], Rainer Claus[3], Harald Binder[2], Michael Lübbert[7,8], Martin Schumacher[2]**

[1]Institute of Genetic Epidemiology, Faculty of Medicine and Medical Center, University of Freiburg, Germany; [2]Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, Germany; [3]Department of Hematology and Oncology, Augsburg University Medical Center, Augsburg, Germany; [4]Department of Internal Medicine III, University Hospital of Ulm, Germany; [5]Division of Cancer Epigenomics, German Cancer Research Center, Heidelberg, Germany; [6]Hematology, Oncology and Tumor Immunology, Campus Virchow Hospital, Charite University Medicine, Berlin, Germany; [7]Department of Hematology-Oncology, Medical Center, Faculty of Medicine, University of Freiburg, Germany; [8]German Consortium for Translational Cancer Research (DKTK), Freiburg, Germany

**Abstract:**

State-of-the art selection methods fail to identify weak but cumulative effects of features found in many high-dimensional omics datasets. Nevertheless, these features play an important role in certain diseases. We present Netboost, a three-step dimension reduction technique. First, a boosting- or Spearman-correlation-based filter is combined with the topological overlap measure to identify the essential edges of the network. Second, sparse hierarchical clustering is applied on the selected edges to identify modules and finally module information is aggregated by the first principal components. We demonstrate the application of the newly developed Netboost in combination with CoxBoost for survival prediction of DNA methylation and gene expression data from 180 acute myeloid leukemia (AML) patients and show, based on cross-validated prediction error curve estimates, its prediction superiority over variable selection on the full dataset as well as over an alternative clustering approach. The identified signature related to chromatin modifying enzymes was replicated in an independent dataset, the phase II AMLSG 12-09 study. In a second application we combine Netboost with Random Forest classification and improve the disease classification error in RNA-sequencing data of Huntington's disease mice. Netboost is a freely available Bioconductor R package for dimension reduction and hypothesis generation in high-dimensional omics applications.

**Contributed Talk**

# Explained Variation in the Linear Mixed Model

**Nicholas Schreck**
DKFZ Heidelberg, Germany

The coefficient of determination is a standard characteristic in linear models with quantitative response variables. It is widely used to assess the proportion of variation explained, to determine the goodness-of-fit and to compare models with different covariates.

However, there has not been an agreement on a similar quantity for the class of linear mixed models yet.

We introduce a natural extension of the well-known adjusted coefficient of determination in linear models to the variance components form of the linear mixed model.

This extension is dimensionless, has an intuitive and simple definition in terms of variance explained, is additive for several random effects and reduces to the adjusted coefficient of determination in the linear model.

To this end, we prove a full decomposition of the sum of squares of the independent variable into the explained and residual variance.

Based on the restricted maximum likelihood equations, we introduce a novel measure for the explained variation which we allocate specifically to the contribution of the fixed and the random covariates of the model.

We illustrate that this empirical explained variation can in particular be used as an improved estimator of the classical additive genetic variance of continuous complex traits.

**Contributed Talk**

# Statistical Methods for Spatial Cluster Detection in Rare Diseases: A Simulation Study of Childhood Cancer Incidence

**Michael Schündeln[1], Toni Lange[2], Maximilian Knoll[3], Claudia Spix[4], Hermann Brenner[5,6,7], Kayan Bozorgmehr[8], Christian Stock[9]**

[1]Pediatric Hematology and Oncology, Department of Pediatrics III, University Hospital Essen and the University of Duisburg-Essen, Essen, Germany.; [2]Center for Evidence-based Healthcare, University Hospital and Faculty of Medicine Carl Gustav Carus, TU Dresden, Germany.; [3]Clinical Cooperation Unit Radiation Oncology, German Cancer Research Center (DKFZ), Heidelberg, Germany.; [4]German Childhood Cancer Registry, Institute for Medical Biostatistics, Epidemiology and Informatics (IMBEI), University Medical Centre of the Johannes Gutenberg University Mainz, Mainz, Germany.; [5]Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany; [6]Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg, Germany; [7]German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany; [8]Department of Population Medicine and Health Services Research, School of Public Health, Bielefeld University, Bielefeld, Germany; [9]Institute of Medical Biometry and Informatics (IMBI), University of Heidelberg, Heidelberg, Germany

**Background and objective:**

The potential existence of spatial clusters in childhood cancer incidence is a debated topic. Identification of such clusters may help to better understand etiology and develop preventive strategies. We evaluated widely used statistical approaches of cluster detection in this context.

**Simulation Study:**

We simulated the incidence of newly diagnosed childhood cancer (140/1,000,000 children under 15 years) and nephroblastoma (7/1,000,000). Clusters of defined size (1 to 50) and relative risk (1 to 100) were randomly assembled on the district level in Germany. For each combination of size and RR 2000 iterations were performed. We then applied three local clustering tests to the simulated data. The Besag-Newell method, the spatial scan statistic and the Bayesian Besag-York-Mollié with Integrated Nested Laplace Approximation approach. We then described the operating characteristics of the tests systematically (such as sensitivity, specificity, predictive values, power etc.).

**Results:**

Depending on the simulated setting, the performance of the tests varied considerably within and between methods. In all methods, the sensitivity was positively associated with increasing size, incidence and RR of the high-risk area. In low RR scenarios, the BYM method showed the highest specificity. In the nephroblastoma scenario compared with the scenario including all cancer cases the performance of all methods was lower.

**Conclusion:**

Reliable inferences on the existence of spatial clusters based on single statistical approaches in childhood cancer remains a challenge. The application of multiple methods, ideally with known operating characteristics, and a critical discussion of the joint evidence is required when aiming to identify high-risk clusters.

**Contributed Talk**

# Variable relation analysis utilizing surrogate variables in random forests

**Stephan Seifert**[1], **Sven Gundlach**[2], **Silke Szymczak**[3]
[1]University of Hamburg; [2]Kiel University; [3]University of Lübeck

The machine learning approach random forests [1] can be successfully applied to omics data, such as gene expression data, for classification or regression. However, the interpretation of the trained prediction models is currently mainly limited to the selection of relevant variables identified based on so-called importance measurements of each individual variable. Thus, relationships between the predictor variables are not considered. We developed a new RF based variable selection method called Surrogate Minimal Depth (SMD) that incorporates variable relations into the selection process of important variables. [2] This is achieved by the exploitation of surrogate variables that have originally been introduced to deal with missing predictor variables. [3] In addition to improving variable selection, surrogate variables and their relationship to the primary split variables measured by the parameter mean adjusted agreement can also be utilized as proxy for the relations between the different variables. This relation analysis goes beyond the investigation of ordinary correlation coefficients because it takes into account the association with the outcome. I will present the basic concept of surrogate variables and mean adjusted agreement, as well as the relation analysis of simulated data as proof of concept and the investigation of experimental breast cancer gene expression datasets to show the practical applicability of this new approach.

**References**

[1] L. Breiman, Mach. Learn. 2001, 45, 5-32.

[2] S. Seifert, S. Gundlach, S. Szymczak, Bioinformatics 2019, 35, 3663-3671.

[3] L. Breiman, J. Friedman, C. J. Stone, R. A. Olshen, Classification and Regression Trees, Taylor & Francis, 1984.

**Contributed Talk**

# What Difference Does Multiple Imputation Make In Longitudinal Modeling of EQ-5D-5L Data: Empirical Analyses of Two Datasets

**Lina Maria Serna Higuita[1], Inka Roesel[1], Fatima Al Sayah[2], Maresa Buchholz[3], Ines Buchholz[3], Thomas Kohlmann[3], Peter Martus[1], You-Shan Feng[1]**

[1]Institute for Clinical Epidemiology and Applied Biostatistics, Medical University of Tübingen, Tübingen, Germany; [2]Alberta PROMs and EQ-5D Research and Support Unit (APERSU), School of Public Health, University of Alberta, Alberta, Canada; [3]Institute for Community Medicine, Medical University Greifswald, Greifswald, Germany

**Background:**

Although multiple imputation (MI) is the state-of-the-art method for managing missing data, it is not clear how missing values in multi-item instruments should be handled, e.g. MI at item or at score level. In addition, longitudinal data analysis techniques such as mixed models (MM) may be equally valid. We therefore explored the differences in modeling the scores of a health-related quality of life questionnaire (EQ-5D-5L) using MM with and without MI at item and score level, in two real data sets.

**Methods:**

We explored 1) Agreement analysis using the observed missing data patterns of EQ-5D-5L responses for a Canadian study, which included patients with type-II diabetes at three time points (Alberta's Caring for Diabetes (ABCD); n=2,040); and 2) Validation analysis using simulated missing patterns for complete cases of a German multi-center study of rehabilitation patients pre- and post-treatment (German Rehabilitation (GR); n=691). Two missing mechanisms (MCAR and MAR) at 8 percentages of missings (5%-65%) were applied to the GR data. Approaches to handle missing EQ-5D-5L scores for all datasets were: Approach-1) MM using respondents with complete cases, approach-2) MM using all available data, approach-3) MM after MI of the EQ-5D-5L scores, and approach-4) MM after MI of EQ-5D-5L items. Agreement was assessed by comparing predicted values and regression coefficients. Validation was examined using mean squared errors (MSE) and standard errors (SE) compared to the original dataset.

**Results:**

**Agreement:**

The ABCD respondents with missing EQ-5D-5L (40.3%) had significantly poorer self-rated health, and lower academic achievement. All 4 approaches estimated similar baseline scores (ABCD≈0.798). At follow up, approach-1 resulted in the highest mean scores (ABCD=0.792) while approach-4 produced the lowest scores (ABCD=0.765). The largest slope of change was observed for approach-4 (visit1–visit3: -0.027), while the smallest slopes were observed for approach-2 (visit3–visit1:-0.011).

**Validation:**

SE and MSE increased with increasing percentages of simulated missing GR data. All approaches showed similar SE and MSE (SE: 0.006-0.011; MSE: 0.032-0.033), however approach-4 showed in the most inaccurate predictions, underestimating the score.

**Discussion:**

In these data, complete case analyses overestimated the scores and MM after MI by items yielded the lowest scores. As there was no loss of accuracy, MM without MI, when baseline covariates are complete, might be the most parsimonious choice to deal with missing data. However, MI may be needed when baseline covariates are missing and/or more than two timepoints are considered.

**Contributed Talk**

# A Machine Learning Approach to Empirical Dynamic Modeling for Biochemical Systems

**Kevin Siswandi**
University of Freiburg, Germany

### Background

In the biosciences, dynamic modeling plays a very important role for understanding and predicting the temporal behaviour of biochemical systems, with wide-ranging applications from bioengineering to precision medicine. Traditionally, dynamic modeling (e.g. in systems biology) is commonly done with Ordinary Differential Equations (ODEs) to predict system dynamics. Such models are typically constructed based on first-principles equations (e.g. Michaelis-Menten kinetics) that are further iteratively modified to be consistent with experiments. Consequently, it could well take several years before a model is quantitatively predictive. Moreover, such ODE models do not scale with increasing amounts of data. At the same time, the demand for high accuracy predictions is increasing in the biotechnology and synthetic biology industry. Here, we investigate a data-driven approach based on machine learning for empirical dynamic modeling that can allow for faster development relative to traditional first-principles modeling, with a particular focus on biochemical systems.

### Methods

We present a numerical framework for a machine learning approach to discover dynamics from time-series data. The main workflow consists of data augmentation, model training and validation, numerical integration, and model explanation. In contrast to other works, our method does not assume any prior (biological) knowledge or governing equations.

Specifically, by posing it as a supervised learning problem, the dynamics can be reconstructed from time-series measurements through solving the resulting optimisation problem. This is done by embedding it within the classical framework of a numerical method (e.g. linear multi-step method or LMM). We evaluate this approach on canonical systems and complex biochemical systems with nonlinear dynamics.

### Results

We show that this method can discover the dynamics of our test systems given enough data. We further find that it could discover bifurcations, is robust to noise, and capable of leveraging additional data to improve its prediction accuracy at scale. Finally, we employ various explainability studies to extract mechanistic insights from the biochemical systems.

### Conclusion

By avoiding assumptions about specific mechanisms, we are able to propose a general machine learning workflow. Thus, it can be applied to any new systems (e.g. pathways or hosts), and could be used to capture complex dynamic relationships which are still unknown in the literature. We believe that it has the potential to accelerate the development of predictive dynamic models due to its data-driven approach.

Poster

# Meta-Cox-regression in DataSHIELD – Federated time-to-event-analysis under data protection constraints

**Ghislain N. Sofack[1], Daniela Zöller[1], Saskia Kiefer[1], Denis Gebele[1], Sebastian Fähndrich[2], Friedrich Kadgien[2], Dennis Hasenpflug[3]**

[1]Institut für Medizinische Biometrie und Statistik (IMBI),Universität Freiburg; [2]Department Innere Medizin, Universitätsklinikum Freiburg; [3]Datenintegrationszentrum, Philipps-Universität Marburg

### Introduction/Background

Studies published so far suggest that Chronic Obstructive Pulmonary Disease (COPD) may be associated with higher rates of mortality in patients with coronavirus disease 2019 (COVID-19). However, the number of cases at a single site is often rather small, making statistical analysis challenging. To address this problem, the data from several sites from the MIRACUM consortium shall be combined. Due to the sensitivity of individual-level data, ethical and practical considerations related to data transmission, and institutional policies, individual-level data cannot be shared. As an alternative, the DataSHIELD framework based on the statistical programming language R can be used. Here, the individual-level data remain within each site and only anonymous aggregated data are shared.

### Problem statement

Up to now, no time-to-event analysis methods are implemented in DataSHIELD. We aim at implementing a meta-regression approach based on the Cox-model in DataSHIELD where only anonymous aggregated data are shared, while simultaneously allowing for explorative, interactive modelling. The approach will be exemplarily applied to explore differences in survival between COVID-19 patients with and those without COPD.

### Methods

Firstly, we present the development of a server-side and client-side DataSHIELD package for calculating survival objects and performing the Cox proportional hazard regression model on individual data at each site. The sensitive patient-level data stored in each server will be processed locally on R studio and only the less-sensitive intermediate statistics like the coefficient's matrices and the Variance Covariance matrices are exchanged and combined via Study Level Meta-Analysis (SLMA) regression techniques to obtain a global analysis. We will demonstrate the process of evaluating the output of the local Cox-regressions for data protection breaches. Exemplarily, we will show the results for comparing the survival of COVID-19 patients with and without COPD using the COVID-19 data distributed across different sites of the MIRACUM consortium.

### Summary

In conclusion, we provide an implementation for SLMA Cox regression in the DataSHIELD framework to enable explorative and interactive modelling for distributed survival data under data protection constraints. We exemplarily demonstrate its applicability to data from the MIRACUM consortium. By demonstrating the process of evaluating the output of the Cox regression for data protection breaches, we rise awareness for the problem.

**Contributed Talk**

# Blinded sample size re-estimation in a paired diagnostic study

**Maria Stark, Antonia Zapf**
University Medical Center Hamburg-Eppendorf, Germany

In a paired confirmatory diagnostic accuracy study, a new experimental test is compared within the same patients to an already existing comparator test. The gold standard defines the true disease status. Hence, each patient undergoes three diagnostic procedures. If feasible and ethically acceptable, regulatory agencies prefer this study design to an unpaired design (CHMP, 2009). The initial sample size calculation is based on assumptions about, among others, the prevalence of the disease and the proportion of discordant test results between the experimental and the comparator test (Miettinen, 1968).

To adjust these assumptions during the study period, an adaptive design for a paired confirmatory diagnostic accuracy study is introduced. This adaptive design is used to re-estimate the prevalence and the proportion of discordant test results to finally re-calculate the sample size. It is a blinded adaptive design as the sensitivity and the specificity of the experimental and comparator test are not re-estimated. Due to the blinding, the type I error rates are not inflated.

An example and a simulation study illustrate the adaptive design. The type I error rate, the power and the sample size of the adaptive design are compared to those of a fixed design. Both designs hold the type I error rate. The adaptive design reaches the advertised power. The fixed design can either be over-or underpowered depending on a possibly wrong assumption regarding the sample size calculation.

The adaptive design compensates inefficiencies of the sample size calculation and therefore it supports to reach the desired study aim.

### References

[1] Committee for Medicinal Products for Human Use, Guideline on clinical evaluation of diagnostic agents. Available at: http://www. ema.europa.eu. 2009, 1-19.

[2] O. S. Miettinen, The matched pairs design in the case of all-or-none responses. 24 2 1968, 339-352.

**Contributed Talk**

# New causal criteria for decisions making under fairness constraints

**Mats Stensrud**
Ecole Polytechnique Fédérale de Lausanne, Switzerland

To justify that a decision is fair, causal reasoning is important: we usually evaluate how the decision was made (the causes of the decision) and what would happen if a different was made (the effects of the decision).

Several causal (counterfactual) definitions of fairness have recently been suggested, but these definitions suffer from any of the following caveats: they rely on ill-defined interventions, require identification conditions that are unreasonably strong, can be gamed by decision makers with malicious intentions, or fail to capture arguably reasonable notions of discrimination.

Motivated by the shortcomings of the existing definitions of fairness, we introduce two new causal criteria to prevent discrimination in practice. These criteria can be applied to settings with non-binary and time-varying decisions. We suggest strategies to evaluate whether these criteria hold in observed data and give conditions that allow identification of counterfactual outcomes under new, non-discriminatory decision rules. The interpretation of our criteria is discussed in several examples.

**Contributed Talk**

# Statistical power for cell identity detection in deep generative models

**Martin Treppner**[1,2], **Harald Binder**[1,2]

[1]Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center - University of Freiburg, Germany; [2]Freiburg Center of Data Analysis and Modelling, Mathematical Institute - Faculty of Mathematics and Physics, University of Freiburg, Germany

One of the most common applications of single-cell RNA-sequencing experiments is to discover groups of cells with a similar expression profile in an attempt to define cell identities. The similarity of these expression profiles is typically examined in a low-dimensional latent space, which can be learned by deep generative models such as variational autoencoders (VAEs). However, the quality of representations in VAEs varies greatly depending on the number of cells under study, which is also reflected in the assignment to specific cell identities. We propose a strategy to answer what number of cells is needed so that a pre-specified percentage of the cells in the latent space is well represented.

We train VAEs on a varying number of cells and evaluate the learned representations' quality by use of the estimated log-likelihood lower bound of each cell. The distribution arising from the values of the log-likelihoods are then compared to a permutation-based distribution of log-likelihoods. We generate the permutation-based distribution by randomly drawing a small subset of cells before training the VAE and permuting each gene's expression values among these randomly drawn cells. By doing so, we ensure that the latent representation's overall structure is preserved, and at the same time, we obtain a null distribution for the log-likelihoods. We then compare log-likelihood distributions for different numbers of cells. We also harness the properties of VAEs by artificially increasing the number of samples in small datasets by generating synthetic data and combining them with the original pilot datasets.

We demonstrate performance on varying sizes of subsamples of the Tabula Muris scRNA-seq dataset from the brain of seven mice processed with the SMART-Seq2 protocol. We show that our approach can be used to plan cell numbers for single-cell RNA-seq experiments, which might improve the reliability of downstream analyses such as cell identity detection and inference of developmental trajectories.

**Poster**

# DIFFERENT STATISTICAL STRATEGIES FOR THE ANALYSIS OF IN VIVO ALKALINE COMET ASSAY DATA

**Timur Tug[1], Annette Bitsch[2], Frank Bringezu[3], Steffi Chang[4], Julia Christin Duda[1], Martina Dammann[5], Roland Frötschl[6], Volker Harm[7], Bernd-Wolfgang Igl[8], Marco Jarzombek[13], Rupert Kellner[2], Fabian Kriegel[13], Jasmin Lott[8], Stefan Pfuhler[9], Ulla Plappert-Helbig[10], Markus Schulz[4], Lea Vaas[7], Marie Vasquez[12], Dietmar Zellner[8], Christina Ziemann[2], Verena Ziegler[11], Katja Ickstadt[1]**

[1]Department of Statistics, TU Dortmund University, Dortmund, Germany; [2]Fraunhofer Institute for Toxicology and Experimental Medicine ITEM, Hannover, Germany; [3]Merck KGaA, Biopharma and Non-Clinical Safety, Darmstadt, Germany; [4]ICCR-Roßdorf GmbH, Rossdorf, Germany; [5]BASF SE, Ludwigshafen am Rhein, Germany; [6]Federal Institute for Drugs and Medical Devices (BfArM), Bonn, Germany; [7]Bayer AG, Berlin, Germany; [8]Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach an der Riss, Germany; [9]Procter & Gamble, Cincinnati, Ohio, USA; [10]Lörrach, Germany; [11]Bayer AG, Wuppertal, Germany; [12]Helix3 Inc, Morrisville, NC, USA; [13]NUVISAN ICB GmbH, Preclinical Compound Profiling,Germany

The in vivo alkaline Comet or single cell gel electrophoresis assay is a standard test in genetic toxicology for measuring DNA damage and repair at an individual cell level. It is a sensitive, fast and simple method to detect single or double strand breaks and therefore, a widespread technique used in several regulatory frameworks today. In 2016, several nonclinical statisticians and toxicologists from academia, industry and one regulatory body founded a working group “Statistics” within the “Gesellschaft für Umwelt-Mutationsforschung e.V.” (GUM). Currently, this interdisciplinary group has collected data from more than 200 experiments performed in various companies to take a closer look on various aspects of the statistical analysis of Comet data.

In this work, we will sketch the assay and related data processing strategies itself. Moreover, we will briefly describe the effect of different summarizing techniques for transferring data from the cell to the slide or animal level, which might influence the final outcome of the test dramatically. Finally, we will present results of various inferential statistical models incl. their comparisons with a special focus on the involvement of historical control data.

**Poster**

# Untersuchung der Qualität der Berichterstattung in RCT Abstracts zu COVID-19 nach CONSORT (CoCo- Studie) – Zwischenbericht eines Reviews

**Sabrina Tulka, Christine Baulig, Stephanie Knippschild**
Lehrstuhl für Medizinische Biometrie und Epidemiologie, Universität Witten/Herdecke, Germany

**Hintergrund:**

Im Jahr 2020 führte die globale COVID-19- Krise aufgrund ihrer Brisanz und Dringlichkeit zu beschleunigter Forschungstätigkeit und Peer-Review-Verfahren. Obwohl die Volltexte zurzeit frei verfügbar sind, sind diese nicht automatisch auch frei zugänglich (z.B. nicht englischsprachig verfasst)! Zusätzlich zwingt ein hoher Zeitdruck medizinisches Personal oftmals dazu, sich ausschließlich über Abstracts einen ersten Überblick in speziellen Themengebieten zu verschaffen. Hierdurch kommt den Abstracts eine Schlüsselrolle zu und bildet nicht selten die Grundlage für Entscheidungen. Das CONSORT-Statement für Abstracts stellt allen Autoren einen Leitfaden zur Verfügung, um die Qualität (Vollständigkeit und Transparenz) der Berichterstattung medizinischer Forschung (auch in Abstracts) zu gewährleisten. Ziel dieser Studie war es die Vollständigkeit in den Abstracts zu allen bisher veröffentlichten COVID-19 RCTs zu untersuchen.

**Methoden:**

Mittels Literaturrecherche in PubMed und Embase wurden alle Publikationen bis zum 29.10.2020 gesucht und hinsichtlich des Themengebietes (berichtet Ergebnisse zu Corona-Studien) und ihres Studiendesigns (RCT) überprüft. Anschließend erfolgte für geeignete Publikationen zum einen die Untersuchung auf Vollständigkeit der Informationen (Information generell aufzufinden) und zum anderen die Prüfung auf Korrektheit (Informationen, gemäß CONSORT für Abstracts berichtet). Grundlage stellte die CONSORT Checkliste für RCT-Abstracts mit insgesamt 16 Items dar. Die Prüfung erfolgte unabhängig durch zwei Bewerter und wurde anschließend konsentiert. Primärer Endpunkt der Studie war der Anteil korrekt umgesetzter CONSORT-Items. Sekundär wurde die Häufigkeit der korrekten Berichterstattung jedes einzelnen Items geprüft.

**Ergebnisse:**

Von insgesamt 88 Publikationen konnten 30 als Veröffentlichung einer RCT in die Analyse eingeschlossen werden. Im Median berichteten die untersuchten Abstracts einen Anteil von 63% der geforderten Kriterien (Quartilspanne: 44% bis 88%, Minimum: 25%, Maximum: 100%). Korrekt umgesetzt wurden im Median 50% der Kriterien (Quartilspanne: 31% bis 70%, Minimum: 12.5%, Maximum: 87.5%). Die „Anzahl der analysierten Patienten“ (20%) und „vollständige Ergebnisse zum primären Endpunkt“ (37%) wurden am seltensten berichtet. Angaben zur Intervention waren in 97% der Abstracts zu finden, aber nur in 43% der Abstracts auch korrekt (vollständig).

**Diskussion:**

Es zeigte sich, dass die Hälfte aller Abstracts maximal die Hälfte aller notwendigen Informationen enthielt. Als besonders auffällig sind hier die Unvollständigkeit zur finalen Patientenzahl, der Ergebnispräsentation sowie zu den jeweils eingesetzten Therapieansätzen für alle Gruppen hervor zu heben. Da (trotz häufiger Verfügbarkeit) in einer schnelllebigen Krisensituation, wie der COVID-19-Pandemie, nur wenig Zeit für eine vollständige und kritische Sichtung aller Volltexte vorhanden ist, müssen Informationen in Studienabstracts vollständig und transparent beschrieben sein. Unsere Untersuchung zeigt, dass ein deutlicher Handlungsbedarf hinsichtlich der Berichtqualität in Abstracts besteht und stellt einen Apell an alle Autoren dar.

**Contributed Talk**

# A Bayesian approach to combine rater assessments

**Lorenz Uhlmann[1,2], Christine Fink[3], Christian Stock[2], Marc Vandemeulebroecke[1], Meinhard Kieser[2]**

[1]Novartis Pharma AG, Basel, Switzerland; [2]Institute of Medical Biometry and Informatics, University of Heidelberg, Heidelberg, Germany; [3]Department of Dermatology, University Medical Center, Ruprecht-Karls University, Heidelberg, Germany

**Background:**

Ideally, endpoints in clinical studies are objectively measurable and easy to assess. However, sometimes this is infeasible and alternative approaches based on (more subjective) rater assessments need to be considered. A Bayesian approach to combine such rater assessments and to estimate relative treatment effects is proposed.

**Methods:**

We focus on a setting where each subject is observed under the condition of every group and where one or multiple raters assign scores that constitute the endpoints. We further assume that the raters compare the arms in a pairwise way by simply scoring them on an individual subject-level. This setting has principle similarities to network meta-analysis where groups (or treatment arms) are ranked in a probabilistic fashion. Many ideas from this field, such as heterogeneity (within raters) or inconsistency (between raters), can be directly applied. We build on Bayesian methodology used in this field and derive models for normally distributed and ordered categorical scores which take into account an arbitrary number of raters and groups.

**Results:**

A general framework is created which is at the same time easy to implement and allows for a straightforward interpretation of the results. The method is illustrated with a real clinical study example on a computer-aided hair detection and removal algorithm in dermatoscopy. Raters assessed the image quality of pictures generated by the algorithm compared to pictures of unshaved and shaved nevis.

**Conclusion:**

A Bayesian approach to combine rater assessments based on an ordinal or continuous scoring system to compare groups in a pairwise fashion is proposed and illustrated using a real data example. The model allows to assess all pairwise comparisons among multiple groups. Since the approach is based on the well-established network meta-analysis methodology, many characteristics can be inferred from that methodology.

**Contributed Talk**

# Mouse clinical trials of N=1: Do we reduce too much?

**Hannes-Friedrich Ulbrich**
Bayer AG, Deutschland

In 2015 the IMI2 7th Call for Proposals requested for 'A comprehensive 'paediatric preclinical POC platform'' for the development of treatments against cancer in children; 'mouse N=1 trials' had to be part of it. The project (ITCC-P4) was launched in 2017.

Four years later the terminology has evolved to 'mouse clinical trials' (MCT). They are experiments where one PDX model (a derivative of a particular patient's tumor) gets implanted into a number of mice to grow and to be treated by different substances: one mouse per substance [and occasionally more for the vehicle -- ITCC-P4 plans with three]. The number of PDX of the same human tumor type is supposed to be "large"; the series of randomized per-patient-tumor experiments are forming a trial. As compared to more 'classical' PDX trials where replicates of mice (usually 6) per substance were used to explore substance differences for one PDX only, mouse clinical trials focus on population response for the considered tumor type. This design is still quite new, "becoming wildly used in pre-clinical oncology drug development, but a statistical framework is yet to be developed" (Guo et al, 2019). Not too much is published yet on whether the reduction to N=1 is reasonable as compared to an imaginable series of 'classical' PDX trials.

Based on data of the already finished OncoTrack IMI project (on colon cancer) we explore the magnitude of differences between the two approaches using resampling techniques.

In this talk we will report the results of this comparison. Statistical models will be described; criteria for comparing these approaches will be discussed.

### References

[1] IMI2 ITCC-P4 Project Description

[2] Guo S et al (2019): Mouse clinical trials in oncology drug development. BMC Cancer 19:718, DOI 10.1186/s12885-019-5907-7

[3] Williams JA (2017) Patient-Derived Xenografts as Cancer Models for Preclinical Drug Screening, DOI 10.1007/978-3-319-55825-7_10

**Contributed Talk**

# Statistical evaluation of the flow cytometric micronucleus in vitro test - same but different

**Lea Al Vaas[1], Robert Smith[2], Jeffrey Bemis[3], Javed Ahmad[2], Steven Bryce[3], Christine Marchand[4], Roland Frötschl[5], Azeddine Elhajouji[6], Ulrike Hemmann[7], Damian McHugh[8], Julia Kenny[9], Natalia Sumption[10], Andreas Zeller[4], Andreas Sutter[10], Daniel Roberts[11]**

[1]Research & Pre-Clinical Statistics Group, Bayer AG, Berlin, Germany; [2]Covance Laboratories Ltd., Harrogate, North Yorkshire, UK; [3]Litron Laboratories, Rochester, NY, USA; [4]Pharmaceutical Sciences, pRED Innovation Center Basel, F. Hoffmann-La Roche Ltd, Basel, Switzerland; [5]Federal Institute for Drugs and Medical Devices (BfArM), Bonn, Germany; [6]Preclinical Safety (PCS), Novartis Institutes for BioMedical Research (NIBR), Basel, Switzerland; [7]Sanofi-Aventis Deutschland GmbH, Frankfurt, Germany; [8]Philip Morris Products S.A., Neuchatel, Switzerland; [9]Genetic Toxicology and Photosafety, GlaxoSmithKline, Ware, Hertfordshire, UK; [10]Bayer AG, Pharmaceuticals, Investigational Toxicology, Berlin, Germany; [11]Genetic and In Vitro Toxicology, Charles River, Skokie, IL, USA

In vitro genotoxicity testing is part of the safety evaluation required for product registration and the initiation of clinical trials. The OECD Test Guideline 487 gives recommendations for the conduct, analysis and interpretation of the in vitro Mammalian Cell Micronucleus (MN) Test. Historically, in vitro MN data have been generated via microscopic examination of cells after exposure to a chemical following scientifically valid, internationally accepted, study designs, that is labour intensive and time consuming. Flow cytometry is an automated technology capable of scoring greater numbers of cells in relatively short time span and analysing genotoxic effects of clastogenic and/or aneugenic origin. However, when acquiring data using flow cytometry, neither the number of cells being evaluated nor the built-in relative survival metrics (cytotoxicity) have undergone critical evaluation for standardization. Herein, we addressed these topics, focusing on the application of the in vitro MN assay scored by flow cytometry (e.g. MicroFlow®) for regulatory purposes. To do so, an international working group comprising genetic toxicologists and statisticians from diverse industry branches, contract research organizations, academia, and regulatory agencies serves as a forum to address the regulatory and technical aspects of submitting GLP-compliant in vitro MN flow cytometry data to support product development and registration.

We will briefly present our motivation and the envisaged initial goals with a focus on the suitability of built-in cytotoxicity metrics for regulatory submissions. Based on a data set collected from multiple cross-industry laboratories the working group additionally evaluates historical control data, recommendations on appropriate study designs, and reviews statistical methods for determining positive micronucleus test results.

**Contributed Talk**

# Acceleration of diagnostic research: Is there a potential for seamless designs?

**Werner Vach[1], Eric Bibiza-Freiwald[2], Oke Gerke[3], Tim Friede[4], Patrick Bossuyt[5], Antonia Zapf[2]**

[1]Basel Academy for Quality and Research in Medicine, Switzerland; [2]Institute of Medical Biometry and Epidemiology, University Medical Center Hamburg-Eppendorf; [3]Department of Nuclear Medicine, Odense University Hospital; [4]Department of Medical Statistics, University Medical Center Goettingen; [5]Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Amsterdam University Medical Centers

**Background:**

New diagnostic tests to identify a well-established disease state have to undergo a series of scientific studies from test construction until finally demonstrating a societal impact. Traditionally, these studies are performed with substantial time gaps in between. Seamless designs allow us to combine a sequence of studies in one protocol and may hence accelerate this process.

**Aim:**

A systematic investigation of the potential of seamless designs in diagnostic research.

**Methods:**

We summarized the major study types in diagnostic research and identified their basic characteristics with respect to applying seamless designs. This information was used to identify major hurdles and opportunities for seamless designs.

**Results:**

11 major study types were identified. The following basic characteristics were identified: type of recruitment (case-control vs population-based), application of a reference standard, inclusion of a comparator, paired or unpaired application of a comparator, assessment of patient relevant outcomes, possibility for blinding of test results.

Two basic hurdles could be identified: 1) Accuracy studies are hard to combine with post-accuracy studies, as the first are required to justify the latter and as application of a reference test in outcome studies is a threat to the study's integrity. 2) Questions, which can be clarified by other study designs, should be clarified before performing a randomized diagnostic study.

However, there is a substantial potential for seamless designs since all steps from the construction until the comparison with the current standard can be combined in one protocol. This may include a switch from case-control to population-based recruitment as well as a switch from a single arm study to a comparative accuracy study. In addition, change in management studies can be combined with an outcome study in discordant pairs. Examples from the literature illustrate the feasibility of both approaches.

**Conclusions:**

There is a potential for seamless designs in diagnostic research.

**Reference:**

[1] Vach W, Bibiza E, Gerke O, Bossuyt PM, Friede T, Zapf A (2021). A potential for seamless designs in diagnostic research could be identified. J Clin Epidemiol. 29:51-59. doi: 10.1016/j.jclinepi.2020.09.019.

**Poster**

# Visualizing uncertainty in diagnostic accuracy studies using comparison regions

**Werner Vach**[1], **Maren Eckert**[2]

[1]Basel Academy for Quality and Research in Medicine, Switzerland; [2]Institute of Medical Biometry and Statistics, University of Freiburg, Germany

The analysis of diagnostic accuracy can be often seen as a two-dimensional estimation problem. The interest is in pairs such as sensitivity and specificity, positive and negative predictive value, or positive and negative likelihood ratio. In visualizing the joint uncertainty in the two-parameter estimate, confidence regions are an obvious choice.

However, Eckert and Vach (2020) recently pointed out, that this a suboptimal approach. Two-dimensional confidence regions support the post-hoc testing of point hypotheses, whereas the evaluation of diagnostic accuracy is related to testing hypotheses on linear combination of parameters (Vach et al. 2012). Consequently, Eckert and Vach suggest the use of comparison regions, supporting such post-hoc tests.

In this poster we illustrate the use of comparison regions in visualizing uncertainty using the results of a published paired diagnostic accuracy study (Ng et al 2008) and contrast it with the use of confidence regions. Both LR-test based and Wald-test based regions are considered. The regions are supplemented by (reference) lines that allow judging possible statements about certain weighted averages of the parameters of interest. We consider the change in sensitivity and specificity as well as the change in the relative frequency of true positive and false positive test results. As the prevalence of the disease state of interest is low in this study, the two approaches give very different results.

Finally, we give some recommendation on the use of comparison regions in analysing diagnostic accuracy studies.

**References:**

[1] Eckert M, Vach W. On the use of comparison regions in visualizing stochastic uncertaint in some two-parameter estimation problems. Biometrical Journal. 2020; 62: 598–609.

[2] Vach W, Gerke O, Høilund-Carlsen PF. Three principles to define the success of a diagnostic study could be identified. J Clin Epidemiol. 2012; 65:293-300.

[3] Ng SH, Chan SC, Liao CT, Chang JT, Ko SF, Wang HM, Chin SC, Lin CY, Huang SF, Yen TC. Distant metastases and synchronous second primary tumors in patients with newly diagnosed oropharyngeal and hypopharyngeal carcinomas: evaluation of (18)F-FDG PET and extended-field multi-detector row CT. Neuroradiology 2008; 50:969-79.

**Contributed Talk**

# Education for Statistics in Practice: Development and evaluation of prediction models: pitfalls and solutions

**Ben Van Calster**[1], **Marten van Smeden**[2]

[1]Department of Development and Regeneration, University of Leuven, Leuven, Belgium; [2]Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands

With fast developments in medical statistics, machine learning and artificial intelligence, the current opportunities for making accurate predictions about the future seem nearly endless. In this lecture we will share some experiences from a medical prediction perspective, where prediction modelling has a long history and models have been implemented in patient care with varying success. We will focus on best practices for the development, evaluation and presentation of prediction models, highlight some common pitfalls, present solutions to circumvent bad prediction modelling and discuss some methodological challenges for the future.

## EXTENDED ABSTRACT

Prediction models are developed throughout science. In this session the focus will be on applications in the medical domain, where prediction models have a long history commonly serving either a diagnostic or prognostic purpose. The ultimate goal of such models is to assist in medical decision making by providing accurate predictions for future individuals.

As we anticipate participants to this session are already well versed in fitting statistical models to data, the focus will be on the common pitfalls when developing statistical (learning) and machine learning models with a prediction aim. Our goal is that the participants gain knowledge about the pitfalls of prediction modeling and increase their familiarity with methods providing solutions for these pitfalls.

The sessions will be arranged in sections of 20 to 30 minutes. The following topics will be covered.

***State of the medical prediction modeling art***

This section begins with a small introduction into the history of prediction modeling in medical research. Positive examples will be highlighted and we will draw from the extensive systematic review literature on clinical prediction models. Recent experiences with a living systematic review on COVID-19 related prediction modeling will be discussed.

***Just another prediction model***

For most health conditions prediction models already exist. How does one prevent that prediction modeling project ends up on the large failed and unused model pile? Using the PROGRESS framework, we discuss various prediction modeling goals. Some good modeling practices and the harm of commonly applied modeling methods are illustrated. Finally, we will highlight some recent developments in formalizing prediction goals (predictimands).

***Methods against overfitting***

Overfitting is arguably the biggest enemy of prediction modeling. There is a large literature on shrinkage estimators that aim at preventing overfitting. In this section we will reflect on the history of shrinkage methods (e.g. Stein's estimator & Le Cessie van Houwelingen heuristic shrinkage) and more recent developments (e.g. lasso and ridge regression variants). The advantages and limitations will be discussed.

***Methods for deciding on appropriate sample size***

Rules of thumb have dominated the discussions on sample size for prediction models for decades (e.g. the need for at least 10 events for every predictor considered). The history and limitations of these rules of thumb will be shown. Recently developed sample size criteria for prediction model development and validation will be presented.

***Model performance and validation***

Validation of prediction models goes beyond evaluation of model coefficients and goodness-of-fit tests. Prediction models should give higher risk estimates for events than for non-events (discrimination). Predictions may be used to support clinical decisions, therefore the risks should be accurate (calibration). We will describe various levels at which a model can be calibrated. Further, the performance of the model to classify patients into low vs high risk patients to support decision making can be evaluated. We discuss decision curve analysis, the most well-known tool for utility validation. The link between calibration and utility is explained.

***Heterogeneity over time and place: there is no such thing as a validated model***

We discuss the different levels of validation (apparent, internal, and external), and what they can tell us. However, it is increasingly recognized that one should expect performance to be heterogeneous between different settings/hospitals. This can be taken into account on many levels: we may focus on having clustered (e.g. multicenter data, IPD) datasets for model development and validation, internal-external cross-validation can be used during model development, and cluster-specific performance can be meta-analyzed at validation. If data allow, meta-regression can be used to gain insight in performance heterogeneity. Model updating can be used to adapt a model to a new setting. In addition, populations tend to change over time. This calls for continuous updating strategies.

***Applied example***

We will describe the development and validation of the ADNEX model to diagnose ovarian cancer. Development, validation, target population, meta-regression, validation studies, model updating, and implementation in ultrasound machines.

***Future perspective: machine learning and AI***

Flexible machine learning algorithms have been around for a while. Recently, however, we have observed a strong increase in their use. We discuss challenges for these methods, such as data hungriness, the risk of automation, increasing complexity of model building, the no free lunch idea, and the winner's curse.

**Contributed Talk**

# Efficient, doubly robust estimation of the effect of dose switching for switchers in a randomised clinical trial

**Kelly Van Lancker**[1], **An Vandebosch**[2], **Stijn Vansteelandt**[1,3]

[1]Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium; [2]Janssen R&D, a division of Janssen Pharmaceutica NV, Beerse, Belgium; [3]Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, United Kingdom

The interpretation of intention-to-treat analyses of randomised clinical trials is often hindered as a result of noncompliance and treatment switching. This has recently given rise to a vigorous research activity on the identification and estimation of so-called estimands.

Motivated by an ongoing clinical trial conducted by Janssen Pharmaceutica in which a flexible dosing regimen is compared to placebo, we evaluate how switchers in the treatment arm (i.e., patients who were switched to the higher dose) would have fared had they been kept on the low dose in order to understand whether flexible dosing is potentially beneficial for them. Comparing these patients' responses with those of patients who stayed on the low dose does not likely entail a satisfactory evaluation because the latter patients are usually in a better health condition and the available information is too limited to enable a reliable adjustment. In view of this, we will transport data from a fixed dosing trial that has been conducted concurrently on the same target, albeit not in an identical patient population.

In particular, we will propose a doubly robust estimator, which relies on an outcome model and a propensity score model for the association between study and patient characteristics. The proposed estimator is easy to evaluate, asymptotically unbiased if either model is correctly specified and efficient (under the restricted semi-parametric model where the randomisation probabilities are known and independent of baseline covariates) when both models are correctly specified. Theoretical properties are also evaluated through Monte Carlo simulations and the method will be illustrated based on the motivating example.

**Contributed Talk**

# Future Prevalence of Type 2 Diabetes – A Comparative Analysis of Chronic Disease Projection Methods

**Dina Voeltz**[1], **Thaddäus Tönnies**[2], **Ralph Brinks**[1,2,3], **Annika Hoyer**[1]

[1]Ludwig-Maximilians-Universität München, Germany; [2]Institute for Biometrics and Epidemiology, German Diabetes Center, Leibniz Institute for Diabetes Research at Heinrich-Heine-University Duesseldorf; [3]Hiller Research Unit for Rheumatology Duesseldorf

**Background:**

Precise projections of future chronic disease cases needing pharmaco-intensive treatments are necessary for effective resource allocation and health care planning in response to increasing disease burden.

**Aim:**

To compare different projection methods to estimate the number of people diagnosed with type 2 diabetes (T2D) in Germany in 2040.

**Methods:**

We compare the results of three methods to project the number of people with T2D in Germany 2040. In a relatively simple approach, method 1) combines the sex- and age-specific prevalence of T2D in 2015 with sex- and age-specific population distributions projected by the German Federal Statistical Office (FSO). Methods 2) and 3) additionally account for incidence of T2D and mortality rates using mathematical relations as proposed by the illness-death model for chronic diseases [1]. Therefore, they are more comprehensive than method 1), which likely adds to their results' validity and accuracy. For this purpose, method 2) firstly models the prevalence of T2D employing a partial differential equation (PDE) which incorporates incidence and mortality [2]. This flexible, yet simple PDE used yields is validated in contexts of dementia, amongst others, and is recommended for chronic disease epidemiology. Subsequently, the estimated prevalence is multiplied with the population projection of the FSO [3]. Hence, method 2) uses the projected general mortality of the FSO and the mortality rate ratio of diseased vs. non-diseased people. By contrast, method 3) estimates future mortality of non-diseased and diseased people independently from the projection of the FSO. These estimated future mortality rates function as input for two PDEs to directly project the absolute number of cases. The sex- and age-specific incidence rate for methods 2) and 3) stems from the risk structure compensation (Risikostrukturausgleich, MorbiRSA) which comprises data from about 70 million Germans in the public health insurance. The incidence rate is assumed to remain as in 2015 throughout the overall projection horizon from 2015 to 2040.

**Results:**

Method 1) projects 8.3 million people with diagnosed T2D in Germany in 2040. Compared to 6.9 million people in 2015, this equals an increase by 21%. Methods 2) and 3) project 11.5 million (+65% compared to 2015) and 12.5 million (+85%) T2D patients, respectively.

**Conclusions:**

The methods' results differ substantially. Method 1) accounts for the aging of the German population but is otherwise relatively little comprehensive. Method 2) and 3) additionally consider underlying changes in the incidence and mortality rates affecting disease prevalence.

**Contributed Talk**

# Statistical challenges in Nursing Science – a practical example

**Maja von Cube**[1], **Martin Wolkewitz**[1], **Christiane Kugler**[2]
[1]Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center - University of Freiburg; [2]Medizinische Fakultät der Albert–Ludwigs–Universität Freiburg, Institut für Pflegewissenschaft Klinisch-Theoretisches Institut des Universitätsklinikums

We give a practical example of a study in Nursing Science. The goal of our study is to investigate whether pets increase the quality of life of patients who had an organ transplantation. As these patients are considered to be at higher risk of acquiring infections, clinical practice has restrictions on holding pets after a transplantation. Nonetheless, pets are presumed to facilitate a healthy lifestyle and thus have a positive impact on human health and well being.

In this practical example, we use data from an observational longitudinal follow up study (n=533) in which clinical parameters, including the acquisition of infections as well as quality of life measurements, were assessed. The latter were measured at seven time points with the Hospital Anxiety and Depression Scale (HADS) and the Short Form health survey (SF-36), a patient reported questionnaire with 36 items. Additionally, information on pets is available for a non-random cross-sectional subsample (n=226).

By combining information from the two datasets, we study whether pets increase the quality of life after an organ transplantation using a linear regression model. Moreover, we use time-to-event analysis to estimate the effect of pets on the time to first infection.

This study bears numerous statistical challenges including the study design, confounding, multiple testing, missing values, competing risks and significant differences in survival between the baseline cohort and the cross-sectional sample. We use this study as a practical example to show how statistical considerations can help to minimize the risk of typical biases arising in clinical epidemiology. Yet, rather than proposing sophisticated statistical approaches, we discuss pragmatic solutions.

**Contributed Talk**

# On the assessment of methods to identify influential points in high-dimensional data

**Shuo Wang, Edwin Kipruto, Willi Sauerbrei**
Medical Center - University of Freiburg, Germany

Extreme values and influential points in predictors often strongly affect the results of statistical analyses in low and high-dimensional settings. Many methods to detect such values have been proposed but there is no consensus on advantages and disadvantages as well as guidance for practice. We will present various classes of methods and illustrate their use in several high-dimensional data. First, we consider a simple pre-transformation which is combined with feature ranking lists to identify influential points, concentrating on univariable situations (Boulesteix and Sauerbrei, 2011, DOI: 10.1002/bimj.201000189). The procedure will be extended by checking for influential points in bivariate models and by adding some steps to the multivariable approach.

Second, to increase stability of feature ranking lists, we will use various aggregation approaches to explore for extreme values in features and influential observations. The former incurs the rank changes of a specific feature, while the latter causes a universal ranking change. For the detection of extreme values, we employ the simple pretransformation on data and detect the features whose ranks significantly changed after the transformation. For the detection of influential observations, we consider a combination of leave-one-out and rank comparison to detect the observations causing large rank changes. These methods are applied in several publicly available datasets.

**Contributed Talk**

# Distribution-free estimation of the partial AUC in diagnostic studies

**Maximilian Wechsung**
Charité - Universitätsmedizin Berlin, Germany

The problem of partial area under the curve (pAUC) estimation arises in diagnostic studies in which not the whole receiver operating characteristic (ROC) curve of a diagnostic test with continuous outcome can be evaluated. Typically, the investigator is bound by economical as well as ethical considerations to analyze only that part of the ROC curve which includes true positive rates and false positive rates above and below certain thresholds, respectively. The pAUC is the area under this partial ROC curve. It can be used to evaluate the performance of a diagnostic test with continuous outcome. In our talk, we consider a distribution-free estimator of the pAUC and establish its asymptotic distribution. The results can be used to construct statistical tests to compare the performance of different diagnostic tests.

**Invited Talk**

# Network meta-analysis for components of complex interventions

**Nicky J Welton**
University of Bristol, UK

Meta-analysis is used to combine results from studies identified in a systematic review comparing specific interventions for a given patient population. However, the validity of the pooled estimate from a meta-analysis relies on the study results being similar enough to pool (homogeneity). Heterogeneity in study results can arise for various reasons, including differences in intervention definitions between studies. Network-meta-analysis (NMA) is an extension of meta-analysis that can combine results from studies to estimate relative effects between multiple (2 or more) interventions, where each study compares some (2 or more) of the interventions of interest. NMA can reduce heterogeneity by treating each intervention definition as a distinct intervention. However, if there are many distinct interventions then evidence networks may be sparse or disconnected so that relative effect estimates are imprecise or not possible to estimate at all. Interventions can sometimes be considered to be made up of component parts, such as some complex interventions or combination therapies.

Component network meta-analysis has been proposed for the synthesis of complex interventions that can be considered a sum of component parts. Component NMA is a form of network meta-regression that estimates the effect of the presence of particular components of an intervention. We discuss methods for categorisation of intervention components, before going on to introduce statistical models for the analysis of the relative efficacy of specific components or combinations of components. The methods respect the randomisation in the included trials and allow the analyst to explore whether the component effects are additive, or if there are interactions between them. The full interaction model corresponds to a standard NMA model.

We illustrate the methods with a range of examples including CBT for depression, electronic interventions for smoking cessation, school-based interventions for anxiety and depression, and psychological interventions for patients with coronary heart disease. We discuss the benefits of component NMA for increasing precision and connecting networks of evidence, the data requirements to fit the models, and make recommendations for the design and reporting of future randomised controlled trials of complex interventions that comprise component parts.

**Contributed Talk**

# Robust Covariance Estimation in Multivariate Meta-Regression

**Thilo Welz**
TU Dortmund University, Germany

Univariate Meta-Regression (MR) is an important technique for medical and psychological research and has been deeply researched. Its multivariate counterpart, however, remains less explored. Multivariate MR holds the potential to incorporate the dependency structure of multiple effect measures as opposed to performing multiple univariate analyses. We explore the possibilities for robust estimation of the covariance of the coefficients in our multivariate MR model. More specifically, we extend heteroscedasticity consistent (also called sandwich or HC-type) estimators from the univariate to the multivariate context. These, along with the Knapp-Hartung adjustment, proved useful in previous work (see Viechtbauer (2015) for an analysis of Knapp-Hartung and Welz & Pauly (2020) for HC-estimators in univariate MR). In our simulations we focus on the bivariate case, which is important for incorporating secondary outcomes as in Copas et al. (2018), but higher dimensions are also possible. The validity of the considered robust estimators is evaluated based on the type-I-error and power of statistical tests based on these estimators. We compare our robust estimation approach with a classical (non-robust) procedure. Finally, we highlight some of the numerical and statistical issues we encountered and provide pointers for others wishing to employ these methods in their analyses.

**Contributed Talk**

# Tree-based Identification of Predictive Factors in Randomized Trials using Weibull Regression

**Wiebke Werft[1], Julia Krzykalla[2], Dominic Edelmann[2], Axel Benner[2]**

[1]Hochschule Mannheim University of Applied Sciences, Germany; [2]German Cancer Research Center (DKFZ), Heidelberg, Germany

Keywords: Predictive biomarkers, Effect modification, Random forest, Time-to-event endpoint, Weibull regression

Novel high-throughput technology provides detailed information on the biomedical characteristics of each patient's disease. These biomarkers may qualify as predictive factors that distinguish patients who benefit from a particular treatment from patients who do not. Hence, large numbers of biomarkers need to be tested in order to gain evidence for tailored treatment decisions ("personalized medicine"). Tree-based methods divide patients into subgroups with differential treatment effects in an automated and data-driven way without requiring extensive pre-specification. Most of these methods mainly aim for a precise prediction of the individual treatment effect, thereby ignoring interpretability of the tree/random forest.

We propose a modification of the model-based recursive partitioning (MOB) approach for subgroup analyses (Seibold, Zeileis et al. 2016), the so-called predMOB, that is able to specifically identify predictive factors (Krzykalla, Benner et al. 2020) from a potentially large number of candidate biomarkers. The original predMOB is developed for normally distributed endpoints only. To widen the field of application, particularly for time-to-event endpoints, we enhanced the predMOB to these situations. More specifically, we use Weibull regression for the base model in the nodes of the tree since MOB and predMOB require fully parametrized models. However, the Weibull model includes the shape parameter as a nuisance parameter which has to be fixed to focus on predictive biomarkers only.

The performance of this extension of the predMOB is assessed concerning identification of the predictive factors as well as prediction accuracy of the individual treatment effect and the predictive effects. Using simulation studies, we are able to show that predMOB provides a targeted approach to predictive factors by reducing the erroneous selection of biomarkers that are only prognostic.

Furthermore, we apply our method to a data set of primary biliary cirrhosis (PBC) patients treated with D-penicillamine or placebo in order to compare our results to those obtained by Su et al. 2008. The aim is to identify predictive factors with respect to overall survival. On the whole, similar variables are identified, but the ranking differs.

**References:**

[1] Krzykalla, J., et al. (2020). "Exploratory identification of predictive biomarkers in randomized trials with normal endpoints." Statistics in Medicine 39(7): 923-939.

[2] Seibold, H., et al. (2016). "Model-based recursive partitioning for subgroup analyses." The International Journal of Biostatistics 12(1): 45-63.

[3] Su. X., et al. (2008). "Interaction trees with censored survival data." The International Journal of Biostatistics 4(1).

**Contributed Talk**

# Statistical Inference for Diagnostic Test Accuracy Studies with Multiple Comparisons

**Max Westphal**[1], **Antonia Zapf**[2]

[1]Fraunhofer Institute for Digital Medicine MEVIS, Bremen, Germany; [2]Institute of Medical Biometry and Epidemiology, UKE Hamburg, Hamburg, Germany

Diagnostic accuracy studies are usually designed to assess the sensitivity and specificity of an index test in relation to a reference standard or established comparative test. This so-called co-primary endpoint analysis has recently been extended to the case that multiple index tests are investigated [1]. Such a design is relevant in modern applications where many different (machine-learned) classification rules based on high dimensional data are considered initially as the final model selection can (partially) be based on data from the diagnostic accuracy study.

In this talk, we motivate the according hypothesis problem and propose different multiple test procedures for that matter. Besides classical parametric corrections (Bonferroni, maxT) we also consider Bootstrap approaches and a Bayesian procedure. We will present early findings from a simulation study to compare the (family-wise) error rate and power of all procedures.

A general observation from the simulation study is the wide variability of rejection rates under different (realistic and least-favorable) parameter configurations. We discuss these findings and possible future extensions of our numerical experiments. All methods have been implemented in a new R package which will also be introduced briefly.

**References:**

[1] Westphal, Max, Antonia Zapf, and Werner Brannath. “A multiple testing framework for diagnostic accuracy studies with co-primary endpoints.” arXiv preprint arXiv:1911.02982 (2019).

**Invited Talk**

# RCT versus RWE: Good versus evil or yin and yang?

**Almut G Winterstein**
University of Florida, USA

Clinicians, researchers and policy makers have been raised in a paradigm that places randomized clinical trials on top of a hierarchy of evidence or that dichotomizes study designs into randomized, which is equated to valid, and not randomized, which is equated to invalid or highly dubious. Major efforts to enhance drug safety research infrastructure have shifted our acceptance of observational designs, especially in instances where the adverse event is not anticipated and unrelated to a drug's indication, resulting in limited confounding. Other instances where evidence from non-randomized studies is accepted include situations where randomization is not feasible. The most recent evolution of real-world evidence as main source of evidence for approval of new molecular entities or indications further challenges our historic understanding of the hierarchy of evidence and the scientific method.

Through randomization and blinding, comparison groups are largely balanced on both measured and unmeasured factors if the trial has sufficient sample size. Protocol-based outcomes ascertainment ensures unbiased, structured assessments regardless of exposure status or baseline characteristics. Used jointly, RCTs can mitigate both selection and measurement biases and support causal inferences. However, besides the escalating cost of RCTs and other feasibility issues, various problems arise that require supplemental methodological approaches to inform regulatory and clinical decision-making, including poor generalizability resulting in inductive fallacy; limited ability to explore effect modification; and significant delays in evidence generation.

Legislative action to address some of these shortcomings was formalized in the United States in the 21st Century Cures Act from 2016, which is designed to help accelerate medical product development. One central component is the concept of real-world evidence, i.e., evidence about the safety and effectiveness of medications derived from real-world data. Importantly, the Cures Act formalizes the concept that valid and actionable evidence can be derived from non-experimental settings using observational study designs and advanced analytic methods.

In this presentation we aim to illustrate that dichotomous approaches that contrast RCTs and RWE are limited in their understanding of the full range of methodological challenges in making causal inferences and then generalizing such inferences for real-world decision-making. Those challenges are discussed across the spectrum of traditional RCTs, pragmatic RCTs that rely on RWD or hybrid designs, and observational studies that rely on RWD. The presentation will end with specific challenges for RWE research in the era of increasing data availability and artificial intelligence.

**Contributed Talk**

# Causal Discovery with Incomplete Cohort Data

**Janine Witte[1,2], Ronja Foraita[1], Ryan M. Andrews[1], Vanessa Didelez[1,2]**
[1]Leibniz Institute for Prevention Research and Epidemiology - BIPS, Bremen, Germany; [2]University of Bremen, Germany

**Background:**

Cohort studies in health research often involve the collection of large numbers of variables over a period of time. They thus form the ideal basis for exploring the relationships among many variables simultaneously, e.g. by methods such as constraint-based causal discovery. These methods aim at inferring a causal graph, combining causal assumptions with statistical tests for conditional independence. A typical problem in practice are missing values. Simple methods for dealing with incomplete data, such as list-wise deletion and mean imputation, can lead to inefficient and biased inference.

**Methods:**

We consider test-wise deletion and multiple imputation for causal discovery. The former applies each conditional independence test to the records containing complete information on the variables used for the test. For multiple imputation, the missing values are imputed M>1 times, the conditional independence test is run on each of the M data sets, and the test results are combined using an appropriate pooling method. We implemented multiple imputation and pooling procedures for causal discovery with continuous, discrete and mixed data. We then compared the performance of test-wise deletion and multiple imputation in scenarios with different missing data patterns typical for cohort data.

**Results:**

Both test-wise deletion and multiple imputation rely on untestable assumptions about the missingness mechanism. Test-wise deletion is computationally simple and can in principle be combined with any conditional independence test. However, it ignores possibly valuable information in partially observed records, hence the power can be low. Multiple imputation has the potential to exploit more information, and outperformed test-wise deletion in several of our simulation scenarios. The simulations also showed, however, that conditional independence testing after multiple imputation is impaired by small sample sizes and large numbers of conditioning variables, especially when the variables are categorical or mixed. Care needs to be taken when choosing the imputation models, as multiple imputation may break down when the number of variables is large, as is typical for cohort studies. Preliminary results suggest that drop-out is best dealt with using test-wise deletion.

**Conclusion:**

Both test-wise deletion and multiple imputation are promising strategies for dealing with missing values in causal discovery, each with their own advantages. Multiple imputation can potentially exploit more information than test-wise deletion, but requires some care when choosing the imputation models. R code for combining test-wise deletion and multiple imputation with different conditional independence tests is available.

**Contributed Talk**

# Pflegewissenschaftliche Versorgungsforschung – Herausforderungen und Chancen

**Karin Wolf-Ostermann**
Universität Bremen, Deutschland

Pflegewissenschaftliche Versorgungsforschung ist einerseits ein Bekenntnis zur Wissenschaftsdisziplin Pflegewissenschaft und andererseits auch ein deutlicher Hinweis darauf, dass sich hieraus auch ein Auftrag zur evidenzbasierten Gestaltung von Versorgung ableitet. Anhand von aktuellen Studien sollen Chancen und Herausforderungen für eine pflegewissenschaftliche Versorgungsforschung am Beispiel Demenz näher beleuchtet werden. Hierbei sollen anhand von Studienbeispielen insbesondere drei Felder näher beleuchtet werden:

1) die Definition von – auch aus Sicht der „Betroffenen“ – relevanten Zielgrößen,

2) die Frage der Zielgruppen von Interventionen

3) die Diskussion passender Studiendesigns– nicht zuletzt mit Blick auf Herausforderungen bei der Evaluation „neuer Technologien“ in der Versorgung.

Hier besteht zukünftig verstärkter Forschungsbedarf in Bezug auf methodische Herausforderungen. Zudem muss stärker als bisher diskutiert werden, wie dem politisch, ethisch und rechtlich fundierten Anliegen der Partizipation Rechnung getragen werden kann. Und nicht zuletzt muss intensiver erörtert werden, wie die Dissemination von Forschungsergebnissen bzw. Implementierung von (evidenzbasierten) Interventionen besser gelingen kann.

**Contributed Talk**

# Genome-wide conditional independence testing with machine learning

**Marvin N. Wright**[1], **David S. Watson**[2,3]

[1]Leibniz Institute for Prevention Research and Epidemiology - BIPS, Bremen, Germany; [2]Oxford Internet Institute, University of Oxford, Oxford, UK; [3]Queen Mary University of London, London, UK

In genetic epidemiology, we are facing extremely high dimensional data and complex patterns such as gene-gene or gene-environment interactions. For this reason, it is promising to use machine learning instead of classical statistical methods to analyze such data. However, most methods for statistical inference with machine learning test against a marginal null hypothesis and by that cannot handle correlated predictor variables.

Building on the knockoff framework of Candès et al. (2018), we propose the conditional predictive impact (CPI), a provably consistent and unbiased estimator of a variables' association with a given outcome, conditional on a reduced set of predictor variables. The method works in conjunction with any supervised learning algorithm and loss function. Simulations confirm that our inference procedures successfully control type I error and achieve nominal coverage probability with greater power than alternative variable importance measures and other nonparametric tests of conditional independence. We apply our method to a gene expression dataset on breast cancer. Further, we propose a modification which avoids the computation of the high-dimensional knockoff matrix and is computationally feasible on data from genome-wide association studies.

**References:**

[1] Candès, E., Fan, Y., Janson, L. and Lv, J. (2018). Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. J Royal Stat Soc Ser B Methodol 80:551–577

# Author Index

## G



## H



## I



## J



## K



## L

## M



## N



## P



## R



## S

## T



## U



## V



## W



## X



## Y



## Z

# Abstract Index by Title

## A



## B



## C



## D

## E



## F



## G



## H



## I



## L

## M



## N



## O



## P



## Q



## R

## S



## T



## U

## V



## W