CENTER FOR MEDICAL DATA SCIENCE
MEDICAL UNIVERSITY OF VIENNA
Institute of Clinical Biometrics

Methodological Publications

2023

1. Akbari N, Heinze G, Rauch G, Sander B, Becher H, Dunkler D: Causal Model Building in the Context of Cardiac Rehabilitation: A Systematic Review. *Int J Environ Res Public Health* (2023) 20(4): 3182; https://doi.org/10.3390/ijerph20043182

   Abstract: *Randomization is an effective design option to prevent bias from confounding in the evaluation of the causal effect of interventions on outcomes. However, in some cases, randomization is not possible, making subsequent adjustment for confounders essential to obtain valid results. Several methods exist to adjust for confounding, with multivariable modeling being among the most widely used. The main challenge is to determine which variables should be included in the causal model and to specify appropriate functional relations for continuous variables in the model. While the statistical literature gives a variety of recommendations on how to build multivariable regression models in practice, this guidance is often unknown to applied researchers. We set out to investigate the current practice of explanatory regression modeling to control confounding in the field of cardiac rehabilitation, for which mainly non-randomized observational studies are available. In particular, we conducted a systematic methods review to identify and compare statistical methodology with respect to statistical model building in the context of the existing recent systematic review CROS-II, which evaluated the prognostic effect of cardiac rehabilitation. CROS-II identified 28 observational studies, which were published between 2004 and 2018. Our methods review revealed that 24 (86%) of the included studies used methods to adjust for confounding. Of these, 11 (46%) mentioned how the variables were selected and two studies (8%) considered functional forms for continuous variables. The use of background knowledge for variable selection was barely reported and data-driven variable selection methods were applied frequently. We conclude that in the majority of studies, the methods used to develop models to investigate the effect of cardiac rehabilitation on outcomes do not meet common criteria for appropriate statistical model building and that reporting often lacks precision.*

2. Geroldinger A, Lusa L, Nold M, Heinze G: Leave-one-out cross-validation, penalization, and differential bias of some prediction model performance measures-a simulation study. *Diagn Progn Res* (2023) 7(1): 9; https://doi.org/10.1186/s41512-023-00146-0

   Abstract: *BACKGROUND: The performance of models for binary outcomes can be described by measures such as the concordance statistic (c-statistic, area under the curve), the discrimination slope, or the Brier score. At internal validation, data resampling techniques, e.g., cross-validation, are frequently employed to correct for optimism in these model performance criteria. Especially with small samples or rare events, leave-one-out cross-validation is a popular choice. METHODS: Using simulations and a real data example, we compared the effect of different resampling techniques on the estimation of c-statistics, discrimination slopes, and Brier scores for three estimators of logistic regression models, including the maximum likelihood and two maximum penalized likelihood estimators. RESULTS: Our simulation study confirms earlier studies reporting that leave-one-out cross-validated c-statistics can be strongly biased towards zero. In addition, our study reveals that this bias is even more pronounced for model estimators shrinking estimated probabilities towards the observed event fraction, such as ridge regression. Leave-one-out cross-validation also provided pessimistic estimates of the discrimination slope but nearly unbiased estimates of the Brier score. CONCLUSIONS: We recommend to use leave-pair-out cross-validation,* *fivefold cross-validation with repetitions, the enhanced or the .632+ bootstrap to estimate c-statistics, and leave-pair-out or fivefold cross-validation to estimate discrimination slopes.*

3. Geroldinger A, Strohmaier S, Kammer M, Schilhart-Wallisch C, Heinze G, Oberbauer R, Haller MC: Sex differences in the survival benefit of kidney transplantation: a retrospective cohort study using target trial emulation. *Nephrol Dial Transplant* (2023) 39(1): 36-44;
https://doi.org/10.1093/ndt/gfad137

   Abstract: *BACKGROUND: Kidney transplantation is the preferred treatment for eligible patients with kidney failure who need renal replacement therapy. However, it remains unclear whether the anticipated survival benefit from kidney transplantation is different for women and men. METHODS: We included all dialysis patients recorded in the Austrian Dialysis and Transplant Registry who were waitlisted for their first kidney transplant between 2000 and 2018. In order to estimate the causal effect of kidney transplantation on 10-year restricted mean survival time, we mimicked a series of controlled clinical trials and applied inverse probability of treatment and censoring weighted sequential Cox models. RESULTS: This study included 4408 patients (33% female) with a mean age of 52 years. Glomerulonephritis was the most common primary renal disease both in women (27%) and men (28%). Kidney transplantation led to a gain of 2.22 years (95% CI 1.88 to 2.49) compared with dialysis over a 10-year follow-up. The effect was smaller in women (1.95 years, 95% CI 1.38 to 2.41) than in men (2.35 years, 95% CI 1.92 to 2.70) due to a better survival on dialysis. Across ages the survival benefit of transplantation over a follow-up of 10 years was smaller in younger women and men and increased with age, showing a peak for both women and men aged about 60 years. CONCLUSIONS: There were few differences in survival benefit by transplantation between females and males. Females had better survival than males on the waitlist receiving dialysis and similar survival to males after transplantation.*

4. Hubin A, Heinze G, De Bin R: Fractional Polynomial Models as Special Cases of Bayesian Generalized Nonlinear Models. *Fractal and Fractional* (2023) 7(9): 641;
https://doi.org/10.3390/fractalfract7090641

   Abstract: *We propose a framework for fitting multivariable fractional polynomial models as specialcases of Bayesian generalized nonlinear models, applying an adapted version of the geneticallymodified mode jumping Markov chain Monte Carlo algorithm. The universality of the Bayesiangeneralized nonlinear models allows us to employ a Bayesian version of fractional polynomials inany supervised learning task, including regression, classification, and time-to-event data analysis.We show through a simulation study that our novel approach performs similarly to the classicalfrequentist multivariable fractional polynomials approach in terms of variable selection, identificationof the true functional forms, and prediction ability, while naturally providing, in contrast to itsfrequentist version, a coherent inference framework. Real-data examples provide further evidence infavor of our approach and show its flexibility.*

5. Janka C, Stamm T, Heinze G, Dorner TE: A Training Programme for Developing Social and Personal Resources and Its Effects on the Perceived Stress Level in Adults in Daily Life-Study Protocol for a Prospective Cohort Study. *Int J Environ Res Public Health* (2023) 20(1): 523;
https://doi.org/10.3390/ijerph20010523

   Abstract: *Persistent stress and insufficient coping strategies have negative consequences for physical and mental health. Teaching adults the skills needed to sustainably improve stress-buffering aspects of their character could contribute to the prevention of stress-related diseases. In this non-randomised, observational, prospective cohort study, participants of a training programme* *for developing social and personal skills, to which they previously self-assigned, are assessed. The 12-month training programme focuses on improving perceived stress level (primary outcome), health behaviour, presence of common somatic symptoms, satisfaction with life, quality of social relationships, and wellbeing by addressing life goals, meaning in life, sense of coherence, social and personal resources, and transcendence. Study participants are recruited from the training groups via the training organiser. Companions, persons with whom they share a close relationship, are recruited to assess the interpersonal diffusion effects of the training. Matched individuals not participating in the training are the control group. Parameter assessment follows a pre-, post-, and follow-up (6 months) design. Designed to improve health-related outcomes in adults by addressing personality characteristics and using Lozanov's superlearning principles to improve learning efficiency, this training programme is, to the study team's knowledge, the first of its kind. From a research perspective, the outcomes of this study can provide new insights into primary prevention of stress-related diseases and how the effects of these measures are passed on through common personal interaction. The trial has been pre-registered (registration number: NCT04165473).*

6. Ma S, Mittlboeck M, Rubio FJ, Liu CC: 2nd special issue on BIOSTATISTICS. *Computational Statistics & Data Analysis* (2023) 181(107681; https://doi.org/10.1016/j.csda.2022.107681

2022

1. Geroldinger A, Blagus R, Ogden H, Heinze G: An investigation of penalization and data augmentation to improve convergence of generalized estimating equations for clustered binary outcomes. *BMC Med Res Methodol* (2022) 22(1): 168; https://doi.org/10.1186/s12874-022-01641-6

   Abstract: *BACKGROUND: In binary logistic regression data are 'separable' if there exists a linear combination of explanatory variables which perfectly predicts the observed outcome, leading to non-existence of some of the maximum likelihood coefficient estimates. A popular solution to obtain finite estimates even with separable data is Firth's logistic regression (FL), which was originally proposed to reduce the bias in coefficient estimates. The question of convergence becomes more involved when analyzing clustered data as frequently encountered in clinical research, e.g. data collected in several study centers or when individuals contribute multiple observations, using marginal logistic regression models fitted by generalized estimating equations (GEE). From our experience we suspect that separable data are a sufficient, but not a necessary condition for non-convergence of GEE. Thus, we expect that generalizations of approaches that can handle separable uncorrelated data may reduce but not fully remove the non-convergence issues of GEE. METHODS: We investigate one recently proposed and two new extensions of FL to GEE. With 'penalized GEE' the GEE are treated as score equations, i.e. as derivatives of a log-likelihood set to zero, which are then modified as in FL. We introduce two approaches motivated by the equivalence of FL and maximum likelihood estimation with iteratively augmented data. Specifically, we consider fully iterated and single-step versions of this 'augmented GEE' approach. We compare the three approaches with respect to convergence behavior, practical applicability and performance using simulated data and a real data example. RESULTS: Our simulations indicate that all three extensions of FL to GEE substantially improve convergence compared to ordinary GEE, while showing a similar or even better performance in terms of accuracy of coefficient estimates and predictions. Penalized GEE often slightly outperforms the augmented GEE approaches, but this comes at the cost of a higher burden of implementation. CONCLUSIONS: When fitting marginal logistic regression models using GEE on sparse data we recommend to apply penalized GEE if one has access to a suitable software implementation and single-step augmented GEE otherwise.*

2. Gregorich M, Melograna F, Sunqvist M, Michiels S, Van Steen K, Heinze G: Individual-specific networks for prediction modelling - A scoping review of methods. *BMC Med Res Methodol* (2022) 22(1): 62; https://doi.org/10.1186/s12874-022-01544-6

   Abstract: *BACKGROUND: Recent advances in biotechnology enable the acquisition of high-dimensional data on individuals, posing challenges for prediction models which traditionally use covariates such as clinical patient characteristics. Alternative forms of covariate representations for the features derived from these modern data modalities should be considered that can utilize their intrinsic interconnection. The connectivity information between these features can be represented as an individual-specific network defined by a set of nodes and edges, the strength of which can vary from individual to individual. Global or local graph-theoretical features describing the network may constitute potential prognostic biomarkers instead of or in addition to traditional covariates and may replace the often unsuccessful search for individual biomarkers in a high-dimensional predictor space. METHODS: We conducted a scoping review to identify, collate and critically appraise the state-of-art in the use of individual-specific networks for prediction modelling in medicine and applied health research, published during 2000-2020 in the electronic databases PubMed, Scopus and Embase. RESULTS: Our scoping review revealed the main application areas* *namely neurology and pathopsychology, followed by cancer research, cardiology and pathology (N = 148). Network construction was mainly based on Pearson correlation coefficients of repeated measurements, but also alternative approaches (e.g. partial correlation, visibility graphs) were found. For covariates measured only once per individual, network construction was mostly based on quantifying an individual's contribution to the overall group-level structure. Despite the multitude of identified methodological approaches for individual-specific network inference, the number of studies that were intended to enable the prediction of clinical outcomes for future individuals was quite limited, and most of the models served as proof of concept that network characteristics can in principle be useful for prediction. CONCLUSION: The current body of research clearly demonstrates the value of individual-specific network analysis for prediction modelling, but it has not yet been considered as a general tool outside the current areas of application. More methodological research is still needed on well-founded strategies for network inference, especially on adequate network sparsification and outcome-guided graph-theoretical feature extraction and selection, and on how networks can be exploited efficiently for prediction modelling.*

3. Hafermann L, Klein N, Rauch G, Kammer M, Heinze G: Using Background Knowledge from Preceding Studies for Building a Random Forest Prediction Model: A Plasmode Simulation Study. *Entropy (Basel)* (2022) 24(6): 847; https://doi.org/10.3390/e24060847

   Abstract: *There is an increasing interest in machine learning (ML) algorithms for predicting patient outcomes, as these methods are designed to automatically discover complex data patterns. For example, the random forest (RF) algorithm is designed to identify relevant predictor variables out of a large set of candidates. In addition, researchers may also use external information for variable selection to improve model interpretability and variable selection accuracy, thereby prediction quality. However, it is unclear to which extent, if at all, RF and ML methods may benefit from external information. In this paper, we examine the usefulness of external information from prior variable selection studies that used traditional statistical modeling approaches such as the Lasso, or suboptimal methods such as univariate selection. We conducted a plasmode simulation study based on subsampling a data set from a pharmacoepidemiologic study with nearly 200,000 individuals, two binary outcomes and 1152 candidate predictor (mainly sparse binary) variables. When the scope of candidate predictors was reduced based on external knowledge RF models achieved better calibration, that is, better agreement of predictions and observed outcome rates. However, prediction quality measured by cross-entropy, AUROC or the Brier score did not improve. We recommend appraising the methodological quality of studies that serve as an external information source for future prediction model development.*

4. Haller MC, Aschauer C, Wallisch C, Leffondre K, van Smeden M, Oberbauer R, Heinze G: Prediction models for living organ transplantation are poorly developed, reported, and validated: a systematic review. *J Clin Epidemiol* (2022) 145:126-135; https://doi.org/10.1016/j.jclinepi.2022.01.025

   Abstract: *OBJECTIVE: To identify and critically appraise risk prediction models for living donor solid organ transplant counselling. STUDY DESIGN AND SETTING: We systematically reviewed articles describing the development or validation of prognostic risk prediction models about living donor solid organ (kidney and liver) transplantation indexed in Medline until April 4, 2021. Models were eligible if intended to predict, at transplant counselling, any outcome occurring after transplantation or donation in recipients or donors. Duplicate study selection, data extraction, assessment for risk of bias and quality of reporting was done using the CHARMS checklist, PRISMA recommendations, PROBAST tool, and TRIPOD Statement. RESULTS: We screened 4691 titles and included 49 studies describing 68 models (35 kidney, 33 liver transplantation). We identified 49 new* *risk prediction models and 19 external validations of existing models. Most models predicted recipients outcomes (n = 38, 75%), e.g., kidney graft loss (29%), or mortality of liver transplant recipients (55%). Many new models (n = 46, 94%) and external validations (n = 17, 89%) had a high risk of bias because of methodological weaknesses. The quality of reporting was generally poor. CONCLUSION: We advise against applying poorly developed, reported, or validated prediction models. Future studies could validate or update the few identified methodologically appropriate models.*

5. Heinze G, Christensen J, Haller MC: Modeling pulse wave velocity trajectories-challenges, opportunities, and pitfalls. *Kidney Int* (2022) 101(3): 459-462; https://doi.org/10.1016/j.kint.2021.12.025

   Abstract: *In this commentary, we discuss the analysis of trajectories of pulse wave velocity in a longitudinal cohort study of children with chronic kidney disease (the Cardiovascular Comorbidity in Children with Chronic Kidney Disease - Transplantation study). We revisit the analysis made by the study authors and unravel some additional limitations. We also reevaluate the implicit assumptions that were made in the chosen analysis and suggest extensions of the basic linear mixed model to obtain more differentiated answers to research questions in nephrology.*

6. Heinze G, van Smeden M, Wynants L, Steyerberg E, van Calster B: Prediction models: stepwise development and simultaneous validation is a step back. *J Clin Epidemiol* (2022) 142:330-331; https://doi.org/10.1016/j.jclinepi.2021.07.019

7. Kammer M, Dunkler D, Michiels S, Heinze G: Evaluating methods for Lasso selective inference in biomedical research: a comparative simulation study. *BMC Med Res Methodol* (2022) 22(1): 206; https://doi.org/10.1186/s12874-022-01681-y

   Abstract: *BACKGROUND: Variable selection for regression models plays a key role in the analysis of biomedical data. However, inference after selection is not covered by classical statistical frequentist theory, which assumes a fixed set of covariates in the model. This leads to over-optimistic selection and replicability issues. METHODS: We compared proposals for selective inference targeting the submodel parameters of the Lasso and its extension, the adaptive Lasso: sample splitting, selective inference conditional on the Lasso selection (SI), and universally valid post-selection inference (PoSI). We studied the properties of the proposed selective confidence intervals available via R software packages using a neutral simulation study inspired by real data commonly seen in biomedical studies. Furthermore, we present an exemplary application of these methods to a publicly available dataset to discuss their practical usability. RESULTS: Frequentist properties of selective confidence intervals by the SI method were generally acceptable, but the claimed selective coverage levels were not attained in all scenarios, in particular with the adaptive Lasso. The actual coverage of the extremely conservative PoSI method exceeded the nominal levels, and this method also required the greatest computational effort. Sample splitting achieved acceptable actual selective coverage levels, but the method is inefficient and leads to less accurate point estimates. The choice of inference method had a large impact on the resulting interval estimates, thereby necessitating that the user is acutely aware of the goal of inference in order to interpret and communicate the results. CONCLUSIONS: Despite violating nominal coverage levels in some scenarios, selective inference conditional on the Lasso selection is our recommended approach for most cases. If simplicity is strongly favoured over efficiency, then sample splitting is an alternative. If only few predictors undergo variable selection (i.e. up to 5) or the avoidance of false positive claims of significance is a concern, then the conservative approach of PoSI may be useful. For the adaptive Lasso, SI should be avoided and only PoSI and sample splitting are recommended. In summary, we find selective* *inference useful to assess the uncertainties in the importance of individual selected predictors for future applications.*

8. LeBlanc M, Rueegg CS, Bekiroglu N, Esterhuizen TM, Fagerland MW, Falk RS, Froslie KF, Graf E, Heinze G, Held U, Holst R, Lange T, Mazumdar M, Myrberg IH, Posch M, Sergeant JC, Vach W, Vance EA, Weedon-Fekjaer H, Zucknick M: Statistical advising: Professional development opportunities for the biostatistician. *Stat Med* (2022) 41(5): 847-859; https://doi.org/10.1002/sim.9290

9. Luijken K, Groenwold RHH, van Smeden M, Strohmaier S, Heinze G: A comparison of full model specification and backward elimination of potential confounders when estimating marginal and conditional causal effects on binary outcomes from observational data. *Biom J* (2022) https://doi.org/10.1002/bimj.202100237

   Abstract: *A common view in epidemiology is that automated confounder selection methods, such as backward elimination, should be avoided as they can lead to biased effect estimates and underestimation of their variance. Nevertheless, backward elimination remains regularly applied. We investigated if and under which conditions causal effect estimation in observational studies can improve by using backward elimination on a prespecified set of potential confounders. An expression was derived that quantifies how variable omission relates to bias and variance of effect estimators. Additionally, 3960 scenarios were defined and investigated by simulations comparing bias and mean squared error (MSE) of the conditional log odds ratio, log(cOR), and the marginal log risk ratio, log(mRR), between full models including all prespecified covariates and backward elimination of these covariates. Applying backward elimination resulted in a mean bias of 0.03 for log(cOR) and 0.02 for log(mRR), compared to 0.56 and 0.52 for log(cOR) and log(mRR), respectively, for a model without any covariate adjustment, and no bias for the full model. In less than 3% of the scenarios considered, the MSE of the log(cOR) or log(mRR) was slightly lower (max 3%) when backward elimination was used compared to the full model. When an initial set of potential confounders can be specified based on background knowledge, there is minimal added value of backward elimination. We advise not to use it and otherwise to provide ample arguments supporting its use.*

10. Mittlbock M, Potschger U, Heinzl H: Weighted pseudo-values for partly unobserved group membership in paediatric stem cell transplantation studies. *Stat Methods Med Res* (2022) 31(1): 76-86; https://doi.org/10.1177/09622802211041756

    Abstract: *Generalised pseudo-values have been suggested to evaluate the impact of allogeneic stem cell transplantation on childhood leukaemia. The approach compares long-term survival of two cohorts defined by the availability or non-availability of suitable donors for stem cell transplantation. A patient's cohort membership becomes known only after completed donor search with or without an identified donor. If a patient suffers an event during donor search, stem cell transplantation will no longer be indicated. In such a case, donor search will be ceased and cohort membership will remain unknown. The generalised pseudo-values approach considers donor identification as binary time-dependent covariate and uses inverse-probability-of-censoring weighting to adjust for non-identified donors. The approach leads to time-consuming computations due to multiple redefinitions of the risk set for pseudo-value calculation and an explicit adjustment for waiting-time bias. Here, the problem is looked at from a different angle. By considering the probability that a donor would have been identified after ceasing of donor search, weights for common pseudo-values are defined. This leads to a faster alternative approach as only a single risk set is necessary. Extensive computer simulations show that both, the generalised and the new weighted pseudo-values approach, provide approximately unbiased estimates. Confidence interval* *coverage is satisfactory for typical clinical scenarios. In situations, where donor identification takes considerably longer than usual, the weighted pseudo-values approach is preferable. Both approaches complement each other as they have different potential in addressing further aspects of the underlying medical question.*

11. van Smeden M, Heinze G, Van Calster B, Asselbergs FW, Vardas PE, Bruining N, de Jaegere P, Moore JH, Denaxas S, Boulesteix AL, Moons KGM: Critical appraisal of artificial intelligence-based prediction models for cardiovascular disease. *Eur Heart J* (2022) 43(31): 2921-2930; https://doi.org/10.1093/eurheartj/ehac238

    Abstract: *The medical field has seen a rapid increase in the development of artificial intelligence (AI)-based prediction models. With the introduction of such AI-based prediction model tools and software in cardiovascular patient care, the cardiovascular researcher and healthcare professional are challenged to understand the opportunities as well as the limitations of the AI-based predictions. In this article, we present 12 critical questions for cardiovascular health professionals to ask when confronted with an AI-based prediction model. We aim to support medical professionals to distinguish the AI-based prediction models that can add value to patient care from the AI that does not.*

12. Wallisch C, Bach P, Hafermann L, Klein N, Sauerbrei W, Steyerberg EW, Heinze G, Rauch G, topic group 2 of the Si: Review of guidance papers on regression modeling in statistical series of medical journals. *PLoS ONE* (2022) 17(1): e0262918; https://doi.org/10.1371/journal.pone.0262918

    Abstract: *Although regression models play a central role in the analysis of medical research projects, there still exist many misconceptions on various aspects of modeling leading to faulty analyses. Indeed, the rapidly developing statistical methodology and its recent advances in regression modeling do not seem to be adequately reflected in many medical publications. This problem of knowledge transfer from statistical research to application was identified by some medical journals, which have published series of statistical tutorials and (shorter) papers mainly addressing medical researchers. The aim of this review was to assess the current level of knowledge with regard to regression modeling contained in such statistical papers. We searched for target series by a request to international statistical experts. We identified 23 series including 57 topic-relevant articles. Within each article, two independent raters analyzed the content by investigating 44 predefined aspects on regression modeling. We assessed to what extent the aspects were explained and if examples, software advices, and recommendations for or against specific methods were given. Most series (21/23) included at least one article on multivariable regression. Logistic regression was the most frequently described regression type (19/23), followed by linear regression (18/23), Cox regression and survival models (12/23) and Poisson regression (3/23). Most general aspects on regression modeling, e.g. model assumptions, reporting and interpretation of regression results, were covered. We did not find many misconceptions or misleading recommendations, but we identified relevant gaps, in particular with respect to addressing nonlinear effects of continuous predictors, model specification and variable selection. Specific recommendations on software were rarely given. Statistical guidance should be developed for nonlinear effects, model specification and variable selection to better support medical researchers who perform or interpret regression analyses.*

2021

1. Frommlet F, Heinze G: Experimental replications in animal trials. *Lab Anim* (2021) 55(1): 65-75; https://doi.org/10.1177/0023677220907617

   Abstract: *The recent discussion on the reproducibility of scientific results is particularly relevant for preclinical research with animal models. Within certain areas of preclinical research, there exists the tradition of repeating an experiment at least twice to demonstrate replicability. If the results of the first two experiments do not agree, then the experiment might be repeated a third time. Sometimes data of one representative experiment are shown; sometimes data from different experiments are pooled. However, there are hardly any guidelines about how to plan for such an experimental design or how to report the results obtained. This article provides a thorough statistical analysis of pre-planned experimental replications as they are currently often applied in practice and gives some recommendations about how to improve on study design and statistical analysis.*

2. Geroldinger A, Hronsky M, Endel F, Endel G, Oberbauer R, Heinze G: Estimation of the prevalence of chronic kidney disease in people with diabetes by combining information from multiple routine data collections. *Journal of the Royal Statistical Society Series a-Statistics in Society* (2021) 184(4): 1260-1282; https://doi.org/10.1111/rssa.12682

   Abstract: *Health care claims databases maintained by social insurance institutions provide rich and sometimes easily accessible data sources for epidemiological research. Interpreting the registered claims, for example, drug prescriptions, as proxies for the condition of interest, for example, diabetes, they allow for nationwide prevalence estimation. We illustrate a more subtle use of health care claims data in estimating the stage-specific prevalence of chronic kidney disease in the Austrian population with diabetes. The main difficulty was that information on the type of disease (chronic or acute) and information on the stage of disease were only available for small, almost disjoint subsets of the health care claims data. Using high-dimensional regression models, we could combine the information and provide nationwide estimates of the stage-specific prevalence of diabetic chronic kidney disease. Validating our estimates by comparing to other studies, we found the level of agreement satisfying.*

3. Gleiss A, Henderson R, Schemper M: Degrees of necessity and of sufficiency: Further results and extensions, with an application to covid-19 mortality in Austria. *Stat Med* (2021) 40(14): 3352-3366; https://doi.org/10.1002/sim.8961

   Abstract: *The purpose of this paper is to extend to ordinal and nominal outcomes the measures of degree of necessity and of sufficiency defined by the authors for dichotomous and survival outcomes in a previous paper. A cause, represented by certain values of prognostic factors, is considered necessary for an event if, without the cause, the event cannot develop. It is considered sufficient for an event if the event is unavoidable in the presence of the cause. The degrees of necessity and sufficiency, ranging from zero to one, are simple, intuitive functions of unconditional and conditional probabilities of an event such as disease or death. These probabilities often will be derived from logistic regression models; the measures, however, do not require any particular model. In addition, we study in detail the relationship between the proposed measures and the related explained variation summary for dichotomous outcomes, which are the common root for the developments for ordinal, nominal, and survival outcomes. We introduce and analyze the Austrian covid-19 data, with the aim of quantifying effects of age and other potentially prognostic factors on covid-19 mortality. This is achieved by standard regression methods but also in terms of the newly proposed measures. It is shown how they complement the toolbox of prognostic factor studies, in particular when comparing the importance of prognostic factors of different types. While the full model's degree of necessity is extremely high (0.933), its low degree of sufficiency (0.179) is responsible for the low proportion of explained variation (0.193).*

4. Gregorich M, Strohmaier S, Dunkler D, Heinze G: Regression with Highly Correlated Predictors: Variable Omission Is Not the Solution. *Int J Environ Res Public Health* (2021) 18(8): 12; https://doi.org/10.3390/ijerph18084259

   Abstract: *Regression models have been in use for decades to explore and quantify the association between a dependent response and several independent variables in environmental sciences, epidemiology and public health. However, researchers often encounter situations in which some independent variables exhibit high bivariate correlation, or may even be collinear. Improper statistical handling of this situation will most certainly generate models of little or no practical use and misleading interpretations. By means of two example studies, we demonstrate how diagnostic tools for collinearity or near-collinearity may fail in guiding the analyst. Instead, the most appropriate way of handling collinearity should be driven by the research question at hand and, in particular, by the distinction between predictive or explanatory aims.*

5. Hafermann L, Becher H, Herrmann C, Klein N, Heinze G, Rauch G: Statistical model building: Background "knowledge" based on inappropriate preselection causes misspecification. *BMC Med Res Methodol* (2021) 21(1): 196; https://doi.org/10.1186/s12874-021-01373-z

   Abstract: *BACKGROUND: Statistical model building requires selection of variables for a model depending on the model's aim. In descriptive and explanatory models, a common recommendation often met in the literature is to include all variables in the model which are assumed or known to be associated with the outcome independent of their identification with data driven selection procedures. An open question is, how reliable this assumed "background knowledge" truly is. In fact, "known" predictors might be findings from preceding studies which may also have employed inappropriate model building strategies. METHODS: We conducted a simulation study assessing the influence of treating variables as "known predictors" in model building when in fact this knowledge resulting from preceding studies might be insufficient. Within randomly generated preceding study data sets, model building with variable selection was conducted. A variable was subsequently considered as a "known" predictor if a predefined number of preceding studies identified it as relevant. RESULTS: Even if several preceding studies identified a variable as a "true" predictor, this classification is often false positive. Moreover, variables not identified might still be truly predictive. This especially holds true if the preceding studies employed inappropriate selection methods such as univariable selection. CONCLUSIONS: The source of "background knowledge" should be evaluated with care. Knowledge generated on preceding studies can cause misspecification.*

6. Sinkovec H, Heinze G, Blagus R, Geroldinger A: To tune or not to tune, a case study of ridge logistic regression in small or sparse datasets. *BMC Med Res Methodol* (2021) 21(1): 199; https://doi.org/10.1186/s12874-021-01374-y

   Abstract: *BACKGROUND: For finite samples with binary outcomes penalized logistic regression such as ridge logistic regression has the potential of achieving smaller mean squared errors (MSE) of coefficients and predictions than maximum likelihood estimation. There is evidence, however, that ridge logistic regression can result in highly variable calibration slopes in small or sparse data situations. METHODS: In this paper, we elaborate this issue further by performing a comprehensive simulation study, investigating the performance of ridge logistic regression in terms of coefficients and predictions and comparing it to Firth's correction that has been shown to perform well in low-dimensional settings. In addition to tuned ridge regression where the penalty strength is estimated from the data by minimizing some measure of the out-of-sample prediction error or information criterion, we also considered ridge regression with pre-specified degree of shrinkage. We included 'oracle' models in the simulation study in which the complexity parameter was chosen based on the true event probabilities (prediction oracle) or regression coefficients (explanation oracle) to demonstrate the capability of ridge regression if truth was known. RESULTS: Performance of ridge regression strongly depends on the choice of complexity parameter. As shown in our simulation and illustrated by a data example, values optimized in small or sparse datasets are negatively correlated with optimal values and suffer from substantial variability which translates into large MSE of coefficients and large variability of calibration slopes. In contrast, in our simulations pre-specifying the degree of shrinkage prior to fitting led to accurate coefficients and predictions even in non-ideal settings such as encountered in the context of rare outcomes or sparse predictors. CONCLUSIONS: Applying tuned ridge regression in small or sparse datasets is problematic as it results in unstable coefficients and predictions. In contrast, determining the degree of shrinkage according to some meaningful prior assumptions about true effects has the potential to reduce bias and stabilize the estimates.*

7. Wallisch C, Agibetov A, Dunkler D, Haller M, Samwald M, Dorffner G, Heinze G: The roles of predictors in cardiovascular risk models - a question of modeling culture? *BMC Med Res Methodol* (2021) 21(1): 284; https://doi.org/10.1186/s12874-021-01487-4

   Abstract: *BACKGROUND: While machine learning (ML) algorithms may predict cardiovascular outcomes more accurately than statistical models, their result is usually not representable by a transparent formula. Hence, it is often unclear how specific values of predictors lead to the predictions. We aimed to demonstrate with graphical tools how predictor-risk relations in cardiovascular risk prediction models fitted by ML algorithms and by statistical approaches may differ, and how sample size affects the stability of the estimated relations. METHODS: We reanalyzed data from a large registry of 1.5 million participants in a national health screening program. Three data analysts developed analytical strategies to predict cardiovascular events within 1 year from health screening. This was done for the full data set and with gradually reduced sample sizes, and each data analyst followed their favorite modeling approach. Predictor-risk relations were visualized by partial dependence and individual conditional expectation plots. RESULTS: When comparing the modeling algorithms, we found some similarities between these visualizations but also occasional divergence. The smaller the sample size, the more the predictor-risk relation depended on the modeling algorithm used, and also sampling variability played an increased role. Predictive performance was similar if the models were derived on the full data set, whereas smaller sample sizes favored simpler models. CONCLUSION: Predictor-risk relations from ML models may differ from those obtained by statistical models, even with large sample sizes. Hence, predictors may assume different roles in risk prediction models. As long as sample size is sufficient, predictive accuracy is not largely affected by the choice of algorithm.*

8. Wallisch C, Dunkler D, Rauch G, de Bin R, Heinze G: Selection of variables for multivariable models: Opportunities and limitations in quantifying model stability by resampling. *Stat Med* (2021) 40(2): 369-381; https://doi.org/10.1002/sim.8779

   Abstract: *Statistical models are often fitted to obtain a concise description of the association of an outcome variable with some covariates. Even if background knowledge is available to guide preselection of covariates, stepwise variable selection is commonly applied to remove irrelevant ones. This practice may introduce additional variability and selection is rarely certain. However, these issues are often ignored and model stability is not questioned. Several resampling-based measures were proposed to describe model stability, including variable inclusion frequencies (VIFs), model selection frequencies, relative conditional bias (RCB), and root mean squared difference ratio (RMSDR). The latter two were recently proposed to assess bias and variance inflation induced by variable selection. Here, we study the consistency and accuracy of resampling estimates of these measures and the optimal choice of the resampling technique. In particular, we compare subsampling and bootstrapping for assessing stability of linear, logistic, and Cox models obtained by backward elimination in a simulation study. Moreover, we exemplify the estimation and interpretation of all suggested measures in a study on cardiovascular risk. The VIF and the model selection frequency are only consistently estimated in the subsampling approach. By contrast, the bootstrap is advantageous in terms of bias and precision for estimating the RCB as well as the RMSDR. Though, unbiased estimation of the latter quantity requires independence of covariates, which is rarely encountered in practice. Our study stresses the importance of addressing model stability after variable selection and shows how to cope with it.*

2020

1. Aalen OO, Stensrud MJ, Didelez V, Daniel R, Roysland K, Strohmaier S: Time-dependent mediators in survival analysis: Modeling direct and indirect effects with the additive hazards model. *Biom J* (2020) 62(3): 532-549; https://doi.org/10.1002/bimj.201800263

   Abstract: *We discuss causal mediation analyses for survival data and propose a new approach based on the additive hazards model. The emphasis is on a dynamic point of view, that is, understanding how the direct and indirect effects develop over time. Hence, importantly, we allow for a time varying mediator. To define direct and indirect effects in such a longitudinal survival setting we take an interventional approach (Didelez, 2018) where treatment is separated into one aspect affecting the mediator and a different aspect affecting survival. In general, this leads to a version of the nonparametric g-formula (Robins, 1986). In the present paper, we demonstrate that combining the g-formula with the additive hazards model and a sequential linear model for the mediator process results in simple and interpretable expressions for direct and indirect effects in terms of relative survival as well as cumulative hazards. Our results generalize and formalize the method of dynamic path analysis (Fosen, Ferkingstad, Borgan, & Aalen, 2006; Strohmaier et al., 2015). An application to data from a clinical trial on blood pressure medication is given.*

2. Bach P, Wallisch C, Klein N, Hafermann L, Sauerbrei W, Steyerberg EW, Heinze G, Rauch G, for topic group 2 of the Si: Systematic review of education and practical guidance on regression modeling for medical researchers who lack a strong statistical background: Study protocol. *PLoS ONE* (2020) 15(12): e0241427; https://doi.org/10.1371/journal.pone.0241427

   Abstract: *In the last decades, statistical methodology has developed rapidly, in particular in the field of regression modeling. Multivariable regression models are applied in almost all medical research projects. Therefore, the potential impact of statistical misconceptions within this field can be enormous Indeed, the current theoretical statistical knowledge is not always adequately transferred to the current practice in medical statistics. Some medical journals have identified this problem and published isolated statistical articles and even whole series thereof. In this systematic review, we aim to assess the current level of education on regression modeling that is provided to medical researchers via series of statistical articles published in medical journals. The present manuscript is a protocol for a systematic review that aims to assess which aspects of regression modeling are covered by statistical series published in medical journals that intend to train and guide applied medical researchers with limited statistical knowledge. Statistical paper series cannot easily be summarized and identified by common keywords in an electronic search engine like Scopus. We therefore identified series by a systematic request to statistical experts who are part or related to the STRATOS Initiative (STRengthening Analytical Thinking for Observational Studies). Within each identified article, two raters will independently check the content of the articles with respect to a predefined list of key aspects related to regression modeling. The content analysis of the topic-relevant articles will be performed using a predefined report form to assess the content as objectively as possible. Any disputes will be resolved by a third reviewer. Summary analyses will identify potential methodological gaps and misconceptions that may have an important impact on the quality of analyses in medical research. This review will thus provide a basis for future guidance papers and tutorials in the field of regression modeling which will enable medical researchers 1) to interpret publications in a correct way, 2) to perform basic statistical analyses in a correct way and 3) to identify situations when the help of a statistical expert is required.*

3. Dunkler D, Haller M, Oberbauer R, Heinze G: To test or to estimate? P-values versus effect sizes. *Transpl Int* (2020) 33(1): 50-55; https://doi.org/10.1111/tri.13535

   Abstract: *Most research in transplant medicine includes statistical analysis of observed data. Too often authors solely rely on P-values derived by statistical tests to answer their research questions. A P-value smaller than 0.05 is typically used to declare "statistical significance" and hence, "proves" that, for example, an intervention has an effect on the outcome of interest. Especially in observational studies, such an approach is highly problematic and can lead to false conclusions. Instead, adequate estimates of the observed size of the effect, for example, expressed as the risk difference, the relative risk or the hazard ratio, should be reported. These effect size measures have to be* *accompanied with an estimate of their precision, like a 95% confidence interval. Such a duo of effect size measure and confidence interval can then be used to answer the important question of clinical relevance.*

4. Sauerbrei W, Perperoglou A, Schmid M, Abrahamowicz M, Becher H, Binder H, Dunkler D, Harrell FE, Jr., Royston P, Heinze G, for TG of the Stratos Initiative: State of the art in selection of variables and functional forms in multivariable analysis-outstanding issues. *Diagn Progn Res* (2020) 4(1): 3;
https://doi.org/10.1186/s41512-020-00074-3

Abstract: *Background: How to select variables and identify functional forms for continuous variables is a key concern when creating a multivariable model. Ad hoc 'traditional' approaches to variable selection have been in use for at least 50 years. Similarly, methods for determining functional forms for continuous variables were first suggested many years ago. More recently, many alternative approaches to address these two challenges have been proposed, but knowledge of their properties and meaningful comparisons between them are scarce. To define a state of the art and to provide evidence-supported guidance to researchers who have only a basic level of statistical knowledge, many outstanding issues in multivariable modelling remain. Our main aims are to identify and illustrate such gaps in the literature and present them at a moderate technical level to the wide community of practitioners, researchers and students of statistics. Methods: We briefly discuss general issues in building descriptive regression models, strategies for variable selection, different ways of choosing functional forms for continuous variables and methods for combining the selection of variables and functions. We discuss two examples, taken from the medical literature, to illustrate problems in the practice of modelling. Results: Our overview revealed that there is not yet enough evidence on which to base recommendations for the selection of variables and functional forms in multivariable analysis. Such evidence may come from comparisons between alternative methods. In particular, we highlight seven important topics that require further investigation and make suggestions for the direction of further research. Conclusions: Selection of variables and of functional forms are important topics in multivariable analysis. To define a state of the art and to provide evidence-supported guidance to researchers who have only a basic level of statistical knowledge, further comparative research is required.*

5. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, Bonten MMJ, Damen JAA, Debray TPA, De Vos M, Dhiman P, Haller MC, Harhay MO, Henckaerts L, Kreuzberger N, Lohman A, Luijken K, Ma J, Andaur CL, Reitsma JB, Sergeant JC, Shi C, Skoetz N, Smits LJM, Snell KIE, Sperrin M, Spijker R, Steyerberg EW, Takada T, van Kuijk SMJ, van Royen FS, Wallisch C, Hooft L, Moons KGM, van Smeden M: Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. *BMJ* (2020) 369:m1328;
https://doi.org/10.1136/bmj.m1328

Abstract: *OBJECTIVE: To review and critically appraise published and preprint reports of prediction models for diagnosing coronavirus disease 2019 (covid-19) in patients with suspected infection, for prognosis of patients with covid-19, and for detecting people in the general population at increased risk of becoming infected with covid-19 or being admitted to hospital with the disease. DESIGN: Living systematic review and critical appraisal by the COVID-PRECISE (Precise Risk Estimation to optimise covid-19 Care for Infected or Suspected patients in diverse sEttings) group. DATA SOURCES: PubMed and Embase through Ovid, arXiv, medRxiv, and bioRxiv up to 5 May 2020. STUDY SELECTION: Studies that developed or validated a multivariable covid-19 related prediction model. DATA EXTRACTION: At least two authors independently extracted data using the CHARMS (critical appraisal and data extraction for systematic reviews of prediction modelling studies) checklist; risk of bias was assessed using PROBAST (prediction model risk of bias assessment tool). RESULTS: 14 217 titles were screened, and 107 studies describing 145 prediction models were included. The review identified four models for identifying people at risk in the general population; 91 diagnostic models for detecting covid-19 (60 were based on medical imaging, nine to diagnose disease severity); and 50 prognostic models for predicting mortality risk, progression to severe disease, intensive care unit admission, ventilation, intubation, or length of hospital stay. The most frequently reported predictors of diagnosis and prognosis of covid-19 are age, body temperature, lymphocyte count, and lung imaging features. Flu-like symptoms and neutrophil count are frequently predictive in diagnostic models, while comorbidities, sex, C reactive protein, and creatinine are frequent prognostic factors. C index estimates ranged from 0.73 to 0.81 in prediction models for the general population, from 0.65 to more than 0.99 in diagnostic models, and from 0.68 to 0.99 in prognostic models. All models were rated at high risk of bias, mostly because of non-representative selection of control patients, exclusion of patients who had not experienced the event of interest by the end of the study, high risk of model overfitting, and vague reporting. Most reports did not include any description of the study population or intended use of the models, and calibration of the model predictions was* *rarely assessed. CONCLUSION: Prediction models for covid-19 are quickly entering the academic literature to support medical decision making at a time when they are urgently needed. This review indicates that proposed models are poorly reported, at high risk of bias, and their reported performance is probably optimistic. Hence, we do not recommend any of these reported prediction models for use in current practice. Immediate sharing of well documented individual participant data from covid-19 studies and collaboration are urgently needed to develop more rigorous prediction models, and validate promising ones. The predictors identified in included models should be considered as candidate predictors for new models. Methodological guidance should be followed because unreliable predictions could cause more harm than benefit in guiding clinical decisions. Finally, studies should adhere to the TRIPOD (transparent reporting of a multivariable prediction model for individual prognosis or diagnosis) reporting guideline. SYSTEMATIC REVIEW REGISTRATION: Protocol https://osf.io/ehc47/, registration https://osf.io/wy245. READERS' NOTE: This article is a living systematic review that will be updated to reflect emerging evidence. Updates may occur for up to two years from the date of original publication. This version is update 2 of the original article published on 7 April 2020 (BMJ 2020;369:m1328), and previous updates can be found as data supplements (https://www.bmj.com/content/369/bmj.m1328/related#datasupp).*

2019

1. Gleiss A, Schemper M: Quantifying degrees of necessity and of sufficiency in cause-effect relationships with dichotomous and survival outcomes. *Stat Med* (2019) 38(23): 4733-4748; https://doi.org/10.1002/sim.8331

   Abstract: *We suggest measures to quantify the degrees of necessity and of sufficiency of prognostic factors for dichotomous and for survival outcomes. A cause, represented by certain values of prognostic factors, is considered necessary for an event if, without the cause, the event cannot develop. It is considered sufficient for an event if the event is unavoidable in the presence of the cause. Necessity and sufficiency can be seen as the two faces of causation, and this symmetry and equal relevance are reflected by the suggested measures. The measures provide an approximate, in some cases an exact, multiplicative decomposition of explained variation as defined by Schemper and Henderson for censored survival and for dichotomous outcomes. The measures, ranging from zero to one, are simple, intuitive functions of unconditional and conditional probabilities of an event such as disease or death. These probabilities often will be derived from logistic or Cox regression models; the measures, however, do not require any particular model. The measures of the degree of necessity implicitly generalize the established attributable fraction or risk for dichotomous prognostic factors and dichotomous outcomes to continuous prognostic factors and to survival outcomes. In a setting with multiple prognostic factors, they provide marginal and partial results akin to marginal and partial odds and hazard ratios from multiple logistic and Cox regression. Properties of the measures are explored by an extensive simulation study. Their application is demonstrated by three typical real data examples.*

2. Heinze G, Wallisch C, Dunkler D: Authors' reply. *Biom J* (2019) 61(6): 1598-1599; https://doi.org/10.1002/bimj.201900196

3. Heinzl H, Benner A: Three brief pieces of statistical advice for medical peer reviewers. *Eur J Clin Invest* (2019) 49(11): e13171; https://doi.org/10.1111/eci.13171

4. Posch M, Bretz F, Friede T, Heinze G: Quantitative approaches underpinning decision making. *Biom J* (2019) 61(5): 1103; https://doi.org/10.1002/bimj.201900202

5. Sinkovec H, Geroldinger A, Heinze G: Bring More Data!-A Good Advice? Removing Separation in Logistic Regression by Increasing Sample Size. *Int J Environ Res Public Health* (2019) 16(23): 4658; https://doi.org/10.3390/ijerph16234658

   Abstract: *The parameters of logistic regression models are usually obtained by the method of maximum likelihood (ML). However, in analyses of small data sets or data sets with unbalanced outcomes or exposures, ML parameter estimates may not exist. This situation has been termed 'separation' as the two outcome groups are separated by the values of a covariate or a linear combination of covariates. To overcome the problem of non-existing ML parameter estimates, applying Firth's correction (FC) was proposed. In practice, however, a principal investigator might be advised to 'bring more data' in order to solve a separation issue. We illustrate the problem by means of examples from colorectal cancer screening and ornithology. It is unclear if such an increasing sample size (ISS) strategy that keeps sampling new observations until separation is removed improves estimation compared to applying FC to the original data set. We performed an extensive simulation study where the main focus was to estimate the cost-adjusted relative efficiency of ML combined with ISS compared to FC. FC yielded reasonably small root mean squared errors and proved to be the more efficient estimator. Given our findings, we propose not to adapt the sample size when separation is encountered but to use FC as the default method of analysis whenever the number of observations or outcome events is critically low.*

2018

1. Dunkler D, Ploner M, Schemper M, Heinze G: Weighted Cox Regression Using the R Package coxphw. *Journal of Statistical Software* (2018) 84(2): 1-26; https://doi.org/10.18637/jss.v084.i02

   Abstract: *Cox's regression model for the analysis of survival data relies on the proportional hazards assumption. However, this assumption is often violated in practice and as a consequence the average relative risk may be under-or overestimated. Weighted estimation of Cox regression is a parsimonious alternative which supplies well interpretable average effects also in case of non-proportional hazards.*
   We provide the R package coxphw implementing weighted Cox regression. By means of two biomedical examples appropriate analyses in the presence of non-proportional hazards are exemplified and advantages of weighted Cox regression are discussed. Moreover, using package coxphw, time-dependent effects can be conveniently estimated by including interactions of covariates with arbitrary functions of time.

2. Gleiss A, Gnant M, Schemper M: Explained variation in shared frailty models. *Stat Med* (2018) 37(9): 1482-1490; https://doi.org/10.1002/sim.7592

   Abstract: *Explained variation measures the relative gain in predictive accuracy when prediction based on prognostic factors replaces unconditional prediction. The factors may be measured on different scales or may be of different types (dichotomous, qualitative, or continuous). Thus, explained variation permits to establish a ranking of the importance of factors, even if predictive accuracy is too low to be helpful in clinical practice. In this contribution, the explained variation measure by Schemper and Henderson (2000) is extended to accommodate random factors, such as center effects in multicenter studies. This permits a direct comparison of the importance of centers and of other prognostic factors. We develop this extension for a shared frailty Cox model and provide an SAS macro and an R function to facilitate its application. Interesting empirical properties of the variation explained by a random factor are explored by a Monte Carlo study. Advantages of the approach are exemplified by an Austrian multicenter study of colon cancer.*

3. Gleiss A, Oberbauer R, Heinze G: An unjustified benefit: immortal time bias in the analysis of time-dependent events. *Transpl Int* (2018) 31(2): 125-130; https://doi.org/10.1111/tri.13081

   Abstract: *Immortal time bias is a problem arising from methodologically wrong analyses of time-dependent events in survival analyses. We illustrate the problem by analysis of a kidney transplantation study. Following patients from transplantation to death, groups defined by the occurrence or nonoccurrence of graft failure during follow-up seemingly had equal overall mortality. Such naive analysis assumes that patients were assigned to the two groups at time of transplantation, which actually are a consequence of occurrence of a time-dependent event later during follow-up. We introduce landmark analysis as the method of choice to avoid immortal time bias. Landmark analysis splits the follow-up time at a common, prespecified time point, the so-called landmark. Groups are then defined by time-dependent events having occurred before the landmark, and outcome events are only considered if occurring after the landmark. Landmark analysis can be easily implemented with common statistical software. In our kidney transplantation example, landmark analyses with landmarks set at 30 and 60 months clearly identified graft failure as a risk factor for overall mortality. We give further typical examples from transplantation research and discuss strengths and limitations of landmark analysis and other methods to address immortal time bias such as Cox regression with time-dependent covariables.*

4. Heinze G, Wallisch C, Dunkler D: Variable selection - A review and recommendations for the practicing statistician. *Biom J* (2018) 60(3): 431-449; https://doi.org/10.1002/bimj.201700067

   Abstract: *Statistical models support medical research by facilitating individualized outcome prognostication conditional on independent variables or by estimating effects of risk factors adjusted for covariates. Theory of statistical models is well-established if the set of independent variables to consider is fixed and small. Hence, we can assume that effect estimates are unbiased and the usual methods for confidence interval estimation are valid. In routine work, however, it is not known a priori which covariates should be included in a model, and often we are confronted with the number of candidate variables in the range 10-30. This number is often too large to be* *considered in a statistical model. We provide an overview of various available variable selection methods that are based on significance or information criteria, penalized likelihood, the change-in-estimate criterion, background knowledge, or combinations thereof. These methods were usually developed in the context of a linear regression model and then transferred to more generalized linear models or models for censored survival data. Variable selection, in particular if used in explanatory modeling where effect estimates are of central interest, can compromise stability of a final model, unbiasedness of regression coefficients, and validity of p-values or confidence intervals. Therefore, we give pragmatic recommendations for the practicing statistician on application of variable selection methods in general (low-dimensional) modeling problems and on performing stability investigations and inference. We also propose some quantities based on resampling the entire variable selection process to be routinely reported by software packages offering automated variable selection algorithms.*

5. Heinzl H, Mittlboeck M: Visualizing the quantile survival time difference curve. *J Eval Clin Pract* (2018) 24(4): 708-712; https://doi.org/10.1111/jep.12948

   Abstract: *The difference between the pth quantiles of 2 survival functions can be used to compare patients' survival between 2 therapies. Setting p = 0.5 yields the median survival time difference. Varying p between 0 and 1 defines the quantile survival time difference curve which can be straightforwardly estimated by the horizontal differences between 2 Kaplan-Meier curves. The estimate's variability can be visualized by adding either a bundle of resampled bootstrap step functions or, alternatively, approximate bootstrap confidence bands. The user-friendly SAS software macro %kmdiff enables the straightforward application of this exploratory graphical approach. The macro is described, and its application is exemplified with breast cancer data. The advantages and limitations of the approach are discussed.*

6. Heinzl H, Mittlboeck M: Contribution to the discussion of "When should meta-analysis avoid making hidden normality assumptions?". *Biom J* (2018) 60(6): 1085-1086; https://doi.org/10.1002/bimj.201800189

7. Mansournia MA, Geroldinger A, Greenland S, Heinze G: Separation in Logistic Regression: Causes, Consequences, and Control. *Am J Epidemiol* (2018) 187(4): 864-870; https://doi.org/10.1093/aje/kwx299

   Abstract: *Separation is encountered in regression models with a discrete outcome (such as logistic regression) where the covariates perfectly predict the outcome. It is most frequent under the same conditions that lead to small-sample and sparse-data bias, such as presence of a rare outcome, rare exposures, highly correlated covariates, or covariates with strong effects. In theory, separation will produce infinite estimates for some coefficients. In practice, however, separation may be unnoticed or mishandled because of software limits in recognizing and handling the problem and in notifying the user. We discuss causes of separation in logistic regression and describe how common software packages deal with it. We then describe methods that remove separation, focusing on the same penalized-likelihood techniques used to address more general sparse-data problems. These methods improve accuracy, avoid software problems, and allow interpretation as Bayesian analyses with weakly informative priors. We discuss likelihood penalties, including some that can be implemented easily with any software package, and their relative advantages and disadvantages. We provide an illustration of ideas and methods using data from a case-control study of contraceptive practices and urinary tract infection.*

8. Meshcheryakova A, Zimmermann P, Ecker R, Mungenast F, Heinze G, Mechtcheriakova D: An Integrative MuSiCO Algorithm: From the Patient-Specific Transcriptional Profiles to Novel Checkpoints in Disease Pathobiology. *Systems Biology* (2018) 351-372; https://doi.org/10.1007/978-3-319-92967-5_18

9. Pötschger U, Heinzl H, Valsecchi MG, Mittlböck M: Assessing the effect of a partly unobserved, exogenous, binary time-dependent covariate on survival probabilities using generalised pseudo-values. *BMC Med Res Methodol* (2018) 18(1): 14; https://doi.org/10.1186/s12874-017-0430-5

   Abstract: *BACKGROUND: Investigating the impact of a time-dependent intervention on the probability of long-term survival is statistically challenging. A typical example is stem-cell transplantation performed after successful donor identification from registered donors. Here, a suggested simple analysis based on the exogenous donor availability status according to registered donors would allow the estimation and comparison of survival probabilities. As donor search is usually ceased after a patient's event, donor availability status is incompletely* *observed, so that this simple comparison is not possible and the waiting time to donor identification needs to be addressed in the analysis to avoid bias. It is methodologically unclear, how to directly address cumulative long-term treatment effects without relying on proportional hazards while avoiding waiting time bias. METHODS: The pseudo-value regression technique is able to handle the first two issues; a novel generalisation of this technique also avoids waiting time bias. Inverse-probability-of-censoring weighting is used to account for the partly unobserved exogenous covariate donor availability. RESULTS: Simulation studies demonstrate unbiasedness and satisfying coverage probabilities of the new method. A real data example demonstrates that study results based on generalised pseudo-values have a clear medical interpretation which supports the clinical decision making process. CONCLUSIONS: The proposed generalisation of the pseudo-value regression technique enables to compare survival probabilities between two independent groups where group membership becomes known over time and remains partly unknown. Hence, cumulative long-term treatment effects are directly addressed without relying on proportional hazards while avoiding waiting time bias.*

2017

1. Gleiss A: Identifiability of Components of Complex Interventions Using Factorial Designs. *J Altern Complement Med* (2017) 23(8): 569-574; https://doi.org/10.1089/acm.2017.0075

   Abstract: *OBJECTIVE: The aim of this contribution is to demonstrate how the component structure of a complex intervention (CI) can be efficiently exploited for study design and statistical analysis by using concepts of factorial designs. Many studies on CIs in complementary and alternative medicine exhibit the structure of factorial designs, where all possible combinations of the levels of two or more treatments occur together. In this contribution, the treatment arms of CI studies are explicitly viewed as factorial combinations of their components. Experimental design offers the general concept of identifiability of effects, that is, unique estimability of the components' effects from the observed data. For factorial designs, a simple cross table representation of the treatment arms can show the components or sums or interactions of components that are identifiable within a given study design. The question of identifiability arises particularly if some combinations of components are not observed (e.g., individualized homeopathic prescription without consultation). Study designs from published homeopathy studies are used for demonstration. CONCLUSIONS: CI studies should explicitly use an intervention's factorial component structure if it is inherent in the treatment arms being compared. In this way, investigators can avoid study designs from which the effects of interest cannot be uniquely estimated and improve the interpretation of estimated effects.*

2. Heinze G, Dunkler D: Five myths about variable selection. *Transpl Int* (2017) 30(1): 6-10; https://doi.org/10.1111/tri.12895

   Abstract: *Multivariable regression models are often used in transplantation research to identify or to confirm baseline variables which have an independent association, causally or only evidenced by statistical correlation, with transplantation outcome. Although sound theory is lacking, variable selection is a popular statistical method which seemingly reduces the complexity of such models. However, in fact, variable selection often complicates analysis as it invalidates common tools of statistical inference such as P-values and confidence intervals. This is a particular problem in transplantation research where sample sizes are often only small to moderate. Furthermore, variable selection requires computer-intensive stability investigations and a particularly cautious interpretation of results. We discuss how five common misconceptions often lead to inappropriate application of variable selection. We emphasize that variable selection and all problems related with it can often be avoided by the use of expert knowledge.*

3. Heinzl H, Mittlboeck M: Assessing a hypothesis test for the difference between two quantiles from independent populations. *COMMUNICATIONS IN STATISTICS-SIMULATION AND COMPUTATION* (2017) 46(5): 3540-3552; https://doi.org/10.1080/03610918.2015.1096379

   Abstract: *An empirical distribution function estimator for the difference of order statistics from two independent populations can be used for inference between quantiles from these populations. The inferential properties of the approach are evaluated in a simulation study where different sample sizes, theoretical distributions, and quantiles are studied. Small to moderate sample sizes, tail quantiles, and quantiles which do not coincide with the expectation of an order statistic are identified as problematic for appropriate Type I error control.*

4. Kabore R, Haller MC, Harambat J, Heinze G, Leffondre K: Risk prediction models for graft failure in kidney transplantation: a systematic review. *Nephrol Dial Transplant* (2017) 32(suppl_2): ii68-ii76; https://doi.org/10.1093/ndt/gfw405

   Abstract: *Risk prediction models are useful for identifying kidney recipients at high risk of graft failure, thus optimizing clinical care. Our objective was to systematically review the models that have been recently developed and validated to predict graft failure in kidney transplantation recipients. We used PubMed and Scopus to search for English, German and French language articles published in 2005-15. We selected studies that developed and validated a new risk prediction model for graft failure after kidney transplantation, or validated an existing model with or without updating the model. Data on recipient characteristics and predictors, as well as modelling and validation methods were extracted. In total, 39 articles met the inclusion criteria. Of these, 34 developed and* *validated a new risk prediction model and 5 validated an existing one with or without updating the model. The most frequently predicted outcome was graft failure, defined as dialysis, re-transplantation or death with functioning graft. Most studies used the Cox model. There was substantial variability in predictors used. In total, 25 studies used predictors measured at transplantation only, and 14 studies used predictors also measured after transplantation. Discrimination performance was reported in 87% of studies, while calibration was reported in 56%. Performance indicators were estimated using both internal and external validation in 13 studies, and using external validation only in 6 studies. Several prediction models for kidney graft failure in adults have been published. Our study highlights the need to better account for competing risks when applicable in such studies, and to adequately account for post-transplant measures of predictors in studies aiming at improving monitoring of kidney transplant recipients.*

5. Lee JW, Lin N, Mittlbock M: Advances in Medical Statistics. *Computational Statistics & Data Analysis* (2017) 113:1-2; https://doi.org/10.1016/j.csda.2017.05.011

6. Puhr R, Heinze G, Nold M, Lusa L, Geroldinger A: Firth's logistic regression with rare events: accurate effect estimates and predictions? *Stat Med* (2017) 36(14): 2302-2317; https://doi.org/10.1002/sim.7273

   Abstract: *Firth's logistic regression has become a standard approach for the analysis of binary outcomes with small samples. Whereas it reduces the bias in maximum likelihood estimates of coefficients, bias towards one-half is introduced in the predicted probabilities. The stronger the imbalance of the outcome, the more severe is the bias in the predicted probabilities. We propose two simple modifications of Firth's logistic regression resulting in unbiased predicted probabilities. The first corrects the predicted probabilities by a post hoc adjustment of the intercept. The other is based on an alternative formulation of Firth's penalization as an iterative data augmentation procedure. Our suggested modification consists in introducing an indicator variable that distinguishes between original and pseudo-observations in the augmented data. In a comprehensive simulation study, these approaches are compared with other attempts to improve predictions based on Firth's penalization and to other published penalization strategies intended for routine use. For instance, we consider a recently suggested compromise between maximum likelihood and Firth's logistic regression. Simulation results are scrutinized with regard to prediction and effect estimation. We find that both our suggested methods do not only give unbiased predicted probabilities but also improve the accuracy conditional on explanatory variables compared with Firth's penalization. While one method results in effect estimates identical to those of Firth's penalization, the other introduces some bias, but this is compensated by a decrease in the mean squared error. Finally, all methods considered are illustrated and compared for a study on arterial closure devices in minimally invasive cardiac surgery. Copyright (c) 2017 John Wiley & Sons, Ltd.*

7. Ternes N, Rotolo F, Heinze G, Michiels S: Identification of biomarker-by-treatment interactions in randomized clinical trials with survival outcomes and high-dimensional spaces. *Biom J* (2017) 59(4): 685-701; https://doi.org/10.1002/bimj.201500234

   Abstract: *Stratified medicine seeks to identify biomarkers or parsimonious gene signatures distinguishing patients that will benefit most from a targeted treatment. We evaluated 12 approaches in high-dimensional Cox models in randomized clinical trials: penalization of the biomarker main effects and biomarker-by-treatment interactions (full-lasso, three kinds of adaptive lasso, ridge+lasso and group-lasso); dimensionality reduction of the main effect matrix via linear combinations (PCA+lasso (where PCA is principal components analysis) or PLS+lasso (where PLS is partial least squares)); penalization of modified covariates or of the arm-specific biomarker effects (two-I model); gradient boosting; and univariate approach with control of multiple testing. We compared these methods via simulations, evaluating their selection abilities in null and alternative scenarios. We varied the number of biomarkers, of nonnull main effects and true biomarker-by-treatment interactions. We also proposed a novel measure evaluating the interaction strength of the developed gene signatures. In the null scenarios, the group-lasso, two-I model, and gradient boosting performed poorly in the presence of nonnull main effects, and performed well in alternative scenarios with also high interaction strength. The adaptive lasso with grouped weights was too conservative. The modified covariates, PCA+lasso, PLS+lasso, and ridge+lasso performed moderately. The full-lasso and adaptive lassos performed well, with the exception of the full-lasso in the presence of only nonnull main effects. The univariate approach performed poorly in alternative scenarios. We also illustrate the methods using gene expression data from 614 breast cancer patients treated with adjuvant chemotherapy.*

2016

1. Dunkler D, Sauerbrei W, Heinze G: Global, Parameterwise and Joint Shrinkage Factor Estimation. *Journal of Statistical Software* (2016) 69(8): 1-19; https://doi.org/10.18637/jss.v069.i08

   Abstract: *The predictive value of a statistical model can often be improved by applying shrinkage methods. This can be achieved, e.g., by regularized regression or empirical Bayes approaches. Various types of shrinkage factors can also be estimated after a maximum likelihood fit has been obtained: while global shrinkage modifies all regression coefficients by the same factor, parameterwise shrinkage factors differ between regression coefficients. The latter ones have been proposed especially in the context of variable selection. With variables which are either highly correlated or associated with regard to contents, such as dummy variables coding a categorical variable, or several parameters describing a nonlinear effect, parameterwise shrinkage factors may not be the best choice. For such cases, we extend the present methodology by so-called 'joint shrinkage factors', a compromise between global and parameterwise shrinkage.*
   Shrinkage factors are often estimated using leave-one-out resampling. We also discuss a computationally simple and much faster approximation to resampling-based shrinkage factor estimation, can be easily obtained in most standard software packages for regression analyses. This alternative may be relevant for simulation studies and other computer-intensive investigations.
   Furthermore, we provide an R package shrink implementing the mentioned shrinkage methods for models fitted by linear, generalized linear, or Cox regression, even if these models involve fractional polynomials or restricted cubic splines to estimate the influence of a continuous variable by a nonlinear function. The approaches and usage of the package shrink are illustrated by means of two examples.

2. Gleiss A, Zeillinger R, Braicu EI, Trillsch F, Vergote I, Schemper M: Statistical controversies in clinical research: the importance of importance. *Ann Oncol* (2016) 27(7): 1185-1189; https://doi.org/10.1093/annonc/mdw159

   Abstract: *We define the notion of 'importance' of prognostic factors in studies of survival and suggest quantifying it by the Schemper-Henderson measure of explained variation. Conceptual differences to the standard approach for the statistical analysis of oncologic studies of survival are discussed and exemplified by means of a study of ovarian cancer. Explained variation permits to establish a ranking of the importance of factors, also if measured on different scales, or of different types (dichotomous, qualitative or continuous), and permits to compare groups of related factors. In practice, the importance of prognostic factors often is disappointingly low. From this, it follows that even strong and highly significant prognostic factors often do not translate into close determination of individual survival of patients.*

3. Heinze G: Statistical reviewing: constructive criticism towards reproducible research. *Transpl Int* (2016) 29(4): 388-389; https://doi.org/10.1111/tri.12758

4. Rinner C, Sauter SK, Endel G, Heinze G, Thurner S, Klimek P, Duftschmid G: Improving the informational continuity of care in diabetes mellitus treatment with a nationwide Shared EHR system: Estimates from Austrian claims data. *Int J Med Inform* (2016) 92(2016): 44-53; https://doi.org/10.1016/j.ijmedinf.2016.05.001

   Abstract: *PURPOSE: Shared Electronic Health Record (EHR) systems, which provide a health information exchange (HIE) within a community of care, were found to be a key enabler of informational continuity of diabetes mellitus (DM) care. Quantitative analyses of the actual contribution of Shared EHR systems to informational continuity of care are rare. The goal of this study was to quantitatively analyze (i) the degree of fragmentation of DM care in Austria as an indicator for the need for HIE, and (ii) the quantity of information (i.e. number of documents) from Austrian DM patients that would be made available by a nationwide Shared EHR system for HIE. METHODS: Our analyses are based on social security claims data of 7.9 million Austrians from 2006 and 2007. DM patients were identified through medication data and inpatient diagnoses. The degree of fragmentation was determined by the number of different healthcare providers per patient. The amount of information that would be made available by a nationwide Shared EHR system was estimated by the number of documents that would have been available to a healthcare provider if he had access to information on the patient's visits to any of the other healthcare providers. As a reference value we determined the number of locally available documents that would have originated from* *the patient's visits to the healthcare provider himself. We performed our analysis for two types of systems: (i) a "comprehensive" Shared EHR system (SEHRS), where each visit of a patient results in a single document (progress note), and (ii) the Austrian ELGA system, which allows four specific document types to be shared. RESULTS: 391,630 DM patients were identified, corresponding to 4.7% of the Austrian population. More than 90% of the patients received health services from more than one healthcare provider in one year. Both, the SEHRS as well as ELGA would have multiplied the available information during a patient visit in comparison to an isolated local EHR system; the median ratio of external to local medical documents was between 1:1 for a typical visit at a primary care provider (SEHRS as well as ELGA) and 39:1 (SEHRS) respectively 28:1 (ELGA) for a typical visit at a hospital. CONCLUSIONS: Due to the high degree of care fragmentation, there is an obvious need for HIE for Austrian DM patients. Both, the SEHRS as well as ELGA could provide a substantial contribution to informational continuity of care in Austrian DM treatment. Hospitals and specialists would have gained the most amount of external information, primary care providers and pharmacies would have at least doubled their available information. Despite being the most important potential feeders of a national Shared EHR system according to our analysis, primary care providers will not tap their full corresponding potential under the current implementation scenario of ELGA.*

2015

1. Gleiss A, Dakna M, Mischak H, Heinze G: Two-group comparisons of zero-inflated intensity values: the choice of test statistic matters. *Bioinformatics* (2015) 31(14): 2310-2317; https://doi.org/10.1093/bioinformatics/btv154

   Abstract: *MOTIVATION: A special characteristic of data from molecular biology is the frequent occurrence of zero intensity values which can arise either by true absence of a compound or by a signal that is below a technical limit of detection. RESULTS: While so-called two-part tests compare mixture distributions between groups, one-part tests treat the zero-inflated distributions as left-censored. The left-inflated mixture model combines these two approaches. Both types of distributional assumptions and combinations of both are considered in a simulation study to compare power and estimation of log fold change. We discuss issues of application using an example from peptidomics.The considered tests generally perform best in scenarios satisfying their respective distributional assumptions. In the absence of distributional assumptions, the two-part Wilcoxon test or the empirical likelihood ratio test is recommended. Assuming a log-normal subdistribution the left-inflated mixture model provides estimates for the proportions of the two considered types of zero intensities. AVAILABILITY: R code is available at http://cemsiis.meduniwien.ac.at/en/kb/science-research/software/*

2. Gobl CS, Bozkurt L, Tura A, Pacini G, Kautzky-Willer A, Mittlbock M: Application of Penalized Regression Techniques in Modelling Insulin Sensitivity by Correlated Metabolic Parameters. *PLoS ONE* (2015) 10(11): e0141524; https://doi.org/10.1371/journal.pone.0141524

   Abstract: *This paper aims to introduce penalized estimation techniques in clinical investigations of diabetes, as well as to assess their possible advantages and limitations. Data from a previous study was used to carry out the simulations to assess: a) which procedure results in the lowest prediction error of the final model in the setting of a large number of predictor variables with high multicollinearity (of importance if insulin sensitivity should be predicted) and b) which procedure achieves the most accurate estimate of regression coefficients in the setting of fewer predictors with small unidirectional effects and moderate correlation between explanatory variables (of importance if the specific relation between an independent variable and insulin sensitivity should be examined). Moreover a special focus is on the correct direction of estimated parameter effects, a non-negligible source of error and misinterpretation of study results. The simulations were performed for varying sample size to evaluate the performance of LASSO, Ridge as well as different algorithms for Elastic Net. These methods were also compared with automatic variable selection procedures (i.e. optimizing AIC or BIC). We were not able to identify one method achieving superior performance in all situations. However, the improved accuracy of estimated effects underlines the importance of using penalized regression techniques in our example (e.g. if a researcher aims to compare relations of several correlated parameters with insulin sensitivity). However, the decision which procedure should be used depends on the specific context of a study (accuracy versus complexity) and moreover should involve clinical prior knowledge.*

3. Kohl M, Plischke M, Leffondre K, Heinze G: PSHREG: a SAS macro for proportional and nonproportional subdistribution hazards regression. *Comput Methods Programs Biomed* (2015) 118(2): 218-233; https://doi.org/10.1016/j.cmpb.2014.11.009

   Abstract: *We present a new SAS macro %pshreg that can be used to fit a proportional subdistribution hazards model for survival data subject to competing risks. Our macro first modifies the input data set appropriately and then applies SAS's standard Cox regression procedure, PROC PHREG, using weights and counting-process style of specifying survival times to the modified data set. The modified data set can also be used to estimate cumulative incidence curves for the event of interest. The application of PROC PHREG has several advantages, e.g., it directly enables the user to apply the Firth correction, which has been proposed as a solution to the problem of undefined (infinite) maximum likelihood estimates in Cox regression, frequently encountered in small sample analyses. Deviation from proportional subdistribution hazards can be detected by both inspecting Schoenfeld-type residuals and testing correlation of these residuals with time, or by including interactions of covariates with functions of time. We illustrate application of these extended methods for competing risk regression using our macro, which is freely available at: http://cemsiis.meduniwien.ac.at/en/kb/science-research/software/statistical-softw are/pshreg,* *by means of analysis of a real chronic kidney disease study. We discuss differences in features and capabilities of %pshreg and the recent (January 2014) SAS PROC PHREG implementation of proportional subdistribution hazards modelling.*

4. Leffondre K, Boucquemont J, Tripepi G, Stel VS, Heinze G, Dunkler D: Analysis of risk factors associated with renal function trajectory over time: a comparison of different statistical approaches. *Nephrol Dial Transplant* (2015) 30(8): 1237-1243; https://doi.org/10.1093/ndt/gfu320

   Abstract: *BACKGROUND: The most commonly used methods to investigate risk factors associated with renal function trajectory over time include linear regression on individual glomerular filtration rate (GFR) slopes, linear mixed models and generalized estimating equations (GEEs). The objective of this study was to explain the principles of these three methods and to discuss their advantages and limitations in particular when renal function trajectories are not completely observable due to dropout. METHODS: We generated data from a hypothetical cohort of 200 patients with chronic kidney disease at inclusion and seven subsequent annual measurements of GFR. The data were generated such that both baseline level and slope of GFR over time were associated with baseline albuminuria status. In a second version of the dataset, we assumed that patients systematically dropped out after a GFR measurement of <15 mL/min/1.73 m(2). Each dataset was analysed with the three methods. RESULTS: The estimated effects of baseline albuminuria status on GFR slope were similar among the three methods when no patient dropped out. When 32.7% dropped out, standard GEE provided biased estimates of the mean GFR slope in normo-, micro- and macroalbuminuric patients. Linear regression on individual slopes and linear mixed models provided slope estimates of the same magnitude, likely because most patients had at least three GFR measurements. However, the linear mixed model was the only method to provide effect estimates on both slope and baseline level of GFR unaffected by dropout. CONCLUSION: This study illustrates that the linear mixed model is the preferred method to investigate risk factors associated with renal function trajectories in studies, where patients may dropout during the study period because of initiation of renal replacement therapy.*

5. Stel VS, Heinze G, Tripepi G, Zoccali C, Jager KJ: Seven essential tools of a cardiologist's survival kit. *Int J Cardiol* (2015) 191(87-89); https://doi.org/10.1016/j.ijcard.2015.04.246

6. Tripepi G, Heinze G, Jager KJ, Stel VS, Dekker FW, Zoccali C: Lag-censoring analysis: lights and shades. *Nephrol Dial Transplant* (2015) 30(5): 700-705; https://doi.org/10.1093/ndt/gfv068

   Abstract: *'Intention-to-treat' (ITT) analysis is the recommended approach for the data analysis of randomized clinical trials (RCT). ITT analysis considers patients in the active or in the control arm as originally allocated by randomization, independently of their actual adherence to the assigned treatment. Lag-censoring analysis is a statistical method which takes into account the compliance of patients to the study protocol because the investigator censors a patient when or shortly after he/she stops the treatment being tested. Herein we describe the methodology underlying lag-censoring analysis in general terms and by considering the application of this technique in the analysis of a large RCT in haemodialysis patients, the Evaluation of Cinacalcet Hydrochloride Therapy to Lower Cardiovascular Events (EVOLVE) trial. Use and misuse of this technique are discussed.*

7. Wakounig S, Heinze G, Schemper M: Non-parametric estimation of relative risk in survival and associated tests. *Stat Methods Med Res* (2015) 24(6): 856-870; https://doi.org/10.1177/0962280211431022

   Abstract: *We extend the Tarone and Ware scheme of weighted log-rank tests to cover the associated weighted Mantel-Haenszel estimators of relative risk. Weighting functions previously employed are critically reviewed. The notion of an average hazard ratio is defined and its connection to the effect size measure P(Y > X) is emphasized. The connection makes estimation of P(Y > X) possible also under censoring. Two members of the extended Tarone-Ware scheme accomplish the estimation of intuitively interpretable average hazard ratios, also under censoring and time-varying relative risk which is achieved by an inverse probability of censoring weighting. The empirical properties of the members of the extended Tarone-Ware scheme are demonstrated by a Monte Carlo study. The differential role of the weighting functions considered is illustrated by a comparative analysis of four real data sets.*

2014

1. Boucquemont J, Heinze G, Jager KJ, Oberbauer R, Leffondre K: Regression methods for investigating risk factors of chronic kidney disease outcomes: the state of the art. *BMC Nephrol* (2014) 15:45; https://doi.org/10.1186/1471-2369-15-45

   Abstract: *BACKGROUND: Chronic kidney disease (CKD) is a progressive and usually irreversible disease. Different types of outcomes are of interest in the course of CKD such as time-to-dialysis, transplantation or decline of the glomerular filtration rate (GFR). Statistical analyses aiming at investigating the association between these outcomes and risk factors raise a number of methodological issues. The objective of this study was to give an overview of these issues and to highlight some statistical methods that can address these topics. METHODS: A literature review of statistical methods published between 2002 and 2012 to investigate risk factors of CKD outcomes was conducted within the Scopus database. The results of the review were used to identify important methodological issues as well as to discuss solutions for each type of CKD outcome. RESULTS: Three hundred and four papers were selected. Time-to-event outcomes were more often investigated than quantitative outcome variables measuring kidney function over time. The most frequently investigated events in survival analyses were all-cause death, initiation of kidney replacement therapy, and progression to a specific value of GFR. While competing risks were commonly accounted for, interval censoring was rarely acknowledged when appropriate despite existing methods. When the outcome of interest was the quantitative decline of kidney function over time, standard linear models focussing on the slope of GFR over time were almost as often used as linear mixed models which allow various numbers of repeated measurements of kidney function per patient. Informative dropout was accounted for in some of these longitudinal analyses. CONCLUSIONS: This study provides a broad overview of the statistical methods used in the last ten years for investigating risk factors of CKD progression, as well as a discussion of their limitations. Some existing potential alternatives that have been proposed in the context of CKD or in other contexts are also highlighted.*

2. Dunkler D, Plischke M, Leffondre K, Heinze G: Augmented backward elimination: a pragmatic and purposeful way to develop statistical models. *PLoS ONE* (2014) 9(11): e113677; https://doi.org/10.1371/journal.pone.0113677

   Abstract: *Statistical models are simple mathematical rules derived from empirical data describing the association between an outcome and several explanatory variables. In a typical modeling situation statistical analysis often involves a large number of potential explanatory variables and frequently only partial subject-matter knowledge is available. Therefore, selecting the most suitable variables for a model in an objective and practical manner is usually a non-trivial task. We briefly revisit the purposeful variable selection procedure suggested by Hosmer and Lemeshow which combines significance and change-in-estimate criteria for variable selection and critically discuss the change-in-estimate criterion. We show that using a significance-based threshold for the change-in-estimate criterion reduces to a simple significance-based selection of variables, as if the change-in-estimate criterion is not considered at all. Various extensions to the purposeful variable selection procedure are suggested. We propose to use backward elimination augmented with a standardized change-in-estimate criterion on the quantity of interest usually reported and interpreted in a model for variable selection. Augmented backward elimination has been implemented in a SAS macro for linear, logistic and Cox proportional hazards regression. The algorithm and its implementation were evaluated by means of a simulation study. Augmented backward elimination tends to select larger models than backward elimination and approximates the unselected model up to negligible differences in point estimates of the regression coefficients. On average, regression coefficients obtained after applying augmented backward elimination were less biased relative to the coefficients of correctly specified models than after backward elimination. In summary, we propose augmented backward elimination as a reproducible variable selection algorithm that gives the analyst more flexibility in adopting model selection to a specific statistical modeling situation.*

3. Leffondre K, Jager KJ, Boucquemont J, Stel VS, Heinze G: Representation of exposures in regression analysis and interpretation of regression coefficients: basic concepts and pitfalls. *Nephrol Dial Transplant* (2014) 29(10): 1806-1814; https://doi.org/10.1093/ndt/gft500

Abstract: *Regression models are being used to quantify the effect of an exposure on an outcome, while adjusting for potential confounders. While the type of regression model to be used is determined by the nature of the outcome variable, e.g. linear regression has to be applied for continuous outcome variables, all regression models can handle any kind of exposure variables. However, some fundamentals of representation of the exposure in a regression model and also some potential pitfalls have to be kept in mind in order to obtain meaningful interpretation of results. The objective of this educational paper was to illustrate these fundamentals and pitfalls, using various multiple regression models applied to data from a hypothetical cohort of 3000 patients with chronic kidney disease. In particular, we illustrate how to represent different types of exposure variables (binary, categorical with two or more categories and continuous), and how to interpret the regression coefficients in linear, logistic and Cox models. We also discuss the linearity assumption in these models, and show how wrongly assuming linearity may produce biased results and how flexible modelling using spline functions may provide better estimates.*

4. Wolbers M, Koller MT, Stel VS, Schaer B, Jager KJ, Leffondre K, Heinze G: Competing risks analyses: objectives and approaches. *Eur Heart J* (2014) 35(42): 2936-2941; https://doi.org/10.1093/eurheartj/ehu131

Abstract: *Studies in cardiology often record the time to multiple disease events such as death, myocardial infarction, or hospitalization. Competing risks methods allow for the analysis of the time to the first observed event and the type of the first event. They are also relevant if the time to a specific event is of primary interest but competing events may preclude its occurrence or greatly alter the chances to observe it. We give a non-technical overview of competing risks concepts for descriptive and regression analyses. For descriptive statistics, the cumulative incidence function is the most important tool. For regression modelling, we introduce regression models for the cumulative incidence function and the cause-specific hazard function, respectively. We stress the importance of choosing statistical methods that are appropriate if competing risks are present. We also clarify the role of competing risks for the analysis of composite endpoints.*

2013

1. Bloching PA, Heinzl H: Assessing the scientific relevance of a single publication over time. *South African Journal of Science* (2013) 109(9-10): 1-2; https://doi.org/10.1590/sajs.2013/20130063

   Abstract: *Quantitatively assessing the scientific relevance of a research paper is challenging for two reasons. Firstly, scientific relevance may change over time, and secondly, it is unclear how to evaluate a recently published paper. The temporally averaged paper-specific impact factor is defined as the yearly average of citations to the paper until now including bonus citations equal to the journal impact factor in the publication year. This new measure subsequently allows relevance rankings and annual updates of all (i.e. both recent and older) scientific papers of a department, or even a whole scientific field, on a more objective basis. It can also be used to assess both the average and overall time-dependent scientific relevance of researchers in a specific department or scientific field.*

2. Heinze G, Ploner M, Beyea J: Confidence intervals after multiple imputation: combining profile likelihood information from logistic regressions. *Stat Med* (2013) 32(29): 5062-5076; https://doi.org/10.1002/sim.5899

   Abstract: *In the logistic regression analysis of a small-sized, case-control study on Alzheimer's disease, some of the risk factors exhibited missing values, motivating the use of multiple imputation. Usually, Rubin's rules (RR) for combining point estimates and variances would then be used to estimate (symmetric) confidence intervals (CIs), on the assumption that the regression coefficients were distributed normally. Yet, rarely is this assumption tested, with or without transformation. In analyses of small, sparse, or nearly separated data sets, such symmetric CI may not be reliable. Thus, RR alternatives have been considered, for example, Bayesian sampling methods, but not yet those that combine profile likelihoods, particularly penalized profile likelihoods, which can remove first order biases and guarantee convergence of parameter estimation. To fill the gap, we consider the combination of penalized likelihood profiles (CLIP) by expressing them as posterior cumulative distribution functions (CDFs) obtained via a chi-squared approximation to the penalized likelihood ratio statistic. CDFs from multiple imputations can then easily be averaged into a combined CDF c , allowing confidence limits for a parameter beta at level 1 - alpha to be identified as those beta* and beta** that satisfy CDF c (beta*) = alpha 2 and CDF c (beta**) = 1 - alpha 2. We demonstrate that the CLIP method outperforms RR in analyzing both simulated data and data from our motivating example. CLIP can also be useful as a confirmatory tool, should it show that the simpler RR are adequate for extended analysis. We also compare the performance of CLIP to Bayesian sampling methods using Markov chain Monte Carlo. CLIP is available in the R package logistf.*

3. Schemper M, Kaider A, Wakounig S, Heinze G: Estimating the correlation of bivariate failure times under censoring. *Stat Med* (2013) 32(27): 4781-4790; https://doi.org/10.1002/sim.5874

   Abstract: *The analysis of correlations within pairs of survival times is of interest to many research topics in medicine, such as the correlation of survival-type endpoints of twins, the correlation of times till failure in paired organs, or the correlation of survival time with a surrogate endpoint. The dependence of such times is assumed monotonic and thus quantification by rank correlation coefficients appropriate. The typical censoring of such times requires more involved methods of estimation and inference as have been developed in recent years. The paper focuses on semiparametric approaches, and in particular on the normal copula-based estimation of Spearman correlation coefficients. The copula approach, often presented for a mathematically inclined readership, is reviewed from the viewpoint of an applied statistician. As an alternative to the maximum likelihood methodology for the normal copula approach (NCE) we introduce an iterative multiple imputation (IMI) method which requires only about 0.05% of the computing time of NCE, without sacrificing statistical performance. For IMI, survival probabilities at death or censoring times are first transformed to normal deviates. Then, those deviates that relate to censored times are iteratively augmented, by using conditional multiple imputation, until convergence is obtained for the normal scores rank correlation, which is similar to Spearman's rank correlation. Statistical properties of NCE and IMI are compared by means of a Monte Carlo study and by means of three real data sets, which also give an impression of the typical range of applications, and of their problems.*

4. Tripepi G, Heinze G, Jager KJ, Stel VS, Dekker FW, Zoccali C: Risk prediction models. *Nephrol Dial Transplant* (2013) 28(8): 1975-1980; https://doi.org/10.1093/ndt/gft095

   Abstract: *Prognostic research focuses on the prediction of the future course of a given disease in probability terms. Prognostication is performed by clinical decision makers by using risk prediction models that allow us to estimate the probability that a specific event occurs in a given patient over a predefined time period conditional on prognostic factors (predictors). Before application in clinical practice, risk prediction models should be properly validated by assessing their discrimination and calibration, or explained variation. Reclassification analyses allow us to evaluate the gain in risk prediction by using a new model compared with an established one. We discuss the concepts of developing and validating risk prediction models by means of two examples, the Framingham risk calculator for prediction of coronary heart disease (CHD), and the recently published Renal Risk Score to predict progression of chronic kidney disease (CKD).*

5. Winkelmayer WC, Heinze G: Assessments of causal effects--theoretically sound, practically unattainable, and clinically not so relevant. *Clin J Am Soc Nephrol* (2013) 8(4): 520-522; https://doi.org/10.2215/CJN.02200213

2012

1. Heinze G: Letter to the editor. *Stat Methods Med Res* (2012) 21(6): 660-661; author reply 665-667; https://doi.org/10.1177/0962280212440533

2. Heinzel A, Fechete R, Söllner J, Perco P, Heinze G, Oberbauer R, Mayer G, Lukas A, Mayer B: Data Graphs for Linking Clinical Phenotype and Molecular Feature Space. *International Journal of Systems Biology and Biomedical Technologies* (2012) 1(1): 11-25; https://doi.org/10.4018/ijsbbt.2012010102

   Abstract: *Omics profiling in translational clinical research has provided detailed molecular characterization of disease phenotypes. Integrating this molecular data space with clinical phenotype descriptors has triggered advancements regarding a systems view on disease, resulting in the concept of stratified medicine. The authors present a methodology for patient stratification by analyzing clinical and molecular information on a per-patient level represented as a data graph. This approach rests on linking patient specific clinical data and biomarker profiles with molecular functional units being derived by segmenting a human proteome interaction network. As a result patient strata are built holding sets of affected functional molecular units as common denominator. Annotation of such functional units on the level of associated diseases, biomarkers and drug targets allows reconciliation with respective clinical data for further improving the assignment of patients to specific strata. The authors finally discuss this approach in the light of adaptive clinical trials design and analysis.*

3. Heinzl H, Waldhoer T: Relevance of the type III error in epidemiological maps. *Int J Health Geogr* (2012) 11(34): 34; https://doi.org/10.1186/1476-072X-11-34

   Abstract: *BACKGROUND: A type III error arises from a two-sided test, when one side is erroneously favoured although the true effect actually resides on the other side. The relevance of this grave error in decision-making is studied for epidemiological maps. RESULTS: Theoretical considerations confirm that a type III error may be large for regions with small numbers of expected cases even when no spatial smoothing has been performed. A simulation study based on infant mortality data in Austria reveals that spatial smoothing may additionally increase the risk of type III errors. CONCLUSIONS: The occurrence of a type III error should be taken into account when interpreting results presented in epidemiological maps, particularly with regard to sparsely populated regions and spatial smoothing.*

4. Mittlbock M, Edler L, LeBlanc M, Niland J, Zwinderman K: Second Issue for Computational Statistics for Clinical Research. *Computational Statistics & Data Analysis* (2012) 56(5): 995-997; https://doi.org/10.1016/j.csda.2012.01.007

2011

1. Dunkler D, Sanchez-Cabo F, Heinze G: Statistical Analysis Principles for Omics Data. *Methods in Molecular Biology (MIMB) 719* (2011) 719(113-131; https://doi.org/10.1007/978-1-61779-027-0_5

   Abstract: *In Omics experiments, typically thousands of hypotheses are tested simultaneously, each based on very few independent replicates. Traditional tests like the t-test were shown to perform poorly with this new type of data. Furthermore, simultaneous consideration of many hypotheses, each prone to a decision error, requires powerful adjustments for this multiple testing situation. After a general introduction to statistical testing, we present the moderated t-statistic, the SAM statistic, and the RankProduct statistic which have been developed to evaluate hypotheses in typical Omics experiments. We also provide an introduction to the multiple testing problem and discuss some state-of-the-art procedures to address this issue. The presented test statistics are subjected to a comparative analysis of a microarray experiment comparing tissue samples of two groups of tumors. All calculations can be done using the freely available statistical software R. Accompanying, commented code is available at: www.meduniwien.ac.at/msi/biometrie/MIMB*

2. Gleiss A, Sanchez-Cabo F, Perco P, Tong D, Heinze G: Adaptive trimmed t-statistics for identifying predominantly high expression in a microarray experiment. *Stat Med* (2011) 30(1): 52-61; https://doi.org/10.1002/sim.4093

   Abstract: *Often, interesting candidate tumor markers are not only genes that show homogeneously higher expression (HHE) in tumor samples compared to control samples, but also genes with only predominantly higher expression (PHE), i.e. genes which exhibit higher expression in at least 80 per cent of tumor samples. Standard parametric test statistics used in the analysis of microarray experiments may fail with PHE as a consequence of the mixture of distributions present in the tumor group. As alternative we consider trimmed t-statistics which compare group mean values after removing outliers in each group. The trimming proportion can be chosen adaptively, either based on a boxplot outlier detection rule or by optimization over a series of tests with varying trimming proportions. The trimmed t-statistics can be plugged into the 'significance analysis of microarrays' (SAM) procedure, yielding the modified boxplot rule test (modBox) and the modified optimization test (modOpt), respectively. By means of simulation of microarray experiments, we show that modOpt is superior to contenders in detecting PHE, while there is only little loss in efficiency under HHE compared to SAM. Analysis of a real microarray experiment revealed that, out of nearly 29 000 genes, about 417 genes exhibiting PHE are detected by modOpt but missed by SAM.*

3. Goliasch G, Blessberger H, Azar D, Heinze G, Wojta J, Bieglmayer C, Wagner O, Schillinger M, Huber K, Maurer G, Haas M, Wiesbauer F: Markers of bone metabolism in premature myocardial infarction (</= 40 years of age). *Bone* (2011) 48(3): 622-626; https://doi.org/10.1016/j.bone.2010.11.005

   Abstract: *INTRODUCTION: Acute myocardial infarction (AMI) at young age is a rare disease with a poor prognosis. Bone metabolism parameters such as 1,25 (OH)(2) vitamin D(3), 25 (OH) vitamin D(3) and osteocalcin have been recently implicated in the development of coronary heart disease (CHD). We evaluated the role of these serum markers in a study population of very young AMI survivors (</= 40 years). METHODS AND RESULTS: We prospectively enrolled 302 subjects into our multi-center case control study, including 102 young myocardial infarction patients (</= 40 years) and 200 control subjects who were frequency-matched on gender and age in an approximate 2:1 ratio per case patient. In the adjusted logistic regression analysis, we used baseline laboratory measurements for the first analysis (acute phase analysis) and measurements from one-year follow-up visits (stable phase analysis). In both, elevated levels of 25 (OH) vitamin D(3) (acute phase: OR per IQR 2.02, 95% CI 1.13-3.58, p = 0.017; stable phase: OR 4.07, 95% CI 1.8-9.21, p = 0.001) and 1,25 (OH)(2) vitamin D(3) (acute phase: OR 2.82, 95% CI 1.7-4.7, p < 0.001; stable phase: OR 4.57, 95% CI 2.31-9.05, p < 0.001) were associated with premature AMI. Conversely, osteocalcin was inversely associated with premature myocardial infarction (acute phase: OR 0.53, 95% CI 0.28-1.03, p = 0.059; stable phase: OR 0.26, 95% CI 0.12-0.6, p < 0.001). The observed associations were independent of the acute phase of myocardial infarction. CONCLUSION: In our study, elevated levels of 25 (OH) vitamin D(3) and 1,25 (OH)(2) vitamin D(3), as well as decreased levels of osteocalcin were* *associated with myocardial infarction in very young patients. The precise mechanism and implications of these findings will have to be elucidated in future studies.*

4. Heinze G: Comment on 'Bias reduction in conditional logistic regression' by J. X. Sun, S. Sinha, S. Wang and T. Maiti, Statistics in Medicine 2010; DOI: 10.1002/sim.4105. *Stat Med* (2011) 30(12): 1466-1467; https://doi.org/10.1002/sim.4173

5. Heinze G, Juni P: An overview of the objectives of and the approaches to propensity score analyses. *Eur Heart J* (2011) 32(14): 1704-1708; https://doi.org/10.1093/eurheartj/ehr031

   Abstract: *The assessment of treatment effects from observational studies may be biased with patients not randomly allocated to the experimental or control group. One way to overcome this conceptual shortcoming in the design of such studies is the use of propensity scores to adjust for differences of the characteristics between patients treated with experimental and control interventions. The propensity score is defined as the probability that a patient received the experimental intervention conditional on pre-treatment characteristics at baseline. Here, we review how propensity scores are estimated and how they can help in adjusting the treatment effect for baseline imbalances. We further discuss how to evaluate adequate overlap of baseline characteristics between patient groups, provide guidelines for variable selection and model building in modelling the propensity score, and review different methods of propensity score adjustments. We conclude that propensity analyses may help in evaluating the comparability of patients in observational studies, and may account for more potential confounding factors than conventional covariate adjustment approaches. However, bias due to unmeasured confounding cannot be corrected for.*

6. Mayer G, Heinze G, Mischak H, Hellemons ME, Lambers Heerspink HJ, Bakker SJL, de Zeeuw D, Haiduk M, Rossing P, Oberbauer R: Omics-Bioinformatics in the Context of Clinical Data. In: *Bioinformatics for Omics Data: Methods and Protocols.* Mayer B (Ed) Humana Press, New York (2011):479-497 https://doi.org/10.1007/978-1-61779-027-0_22

   Abstract: *The Omics revolution has provided the researcher with tools and methodologies for qualitative and quantitative assessment of a wide spectrum of molecular players spanning from the genome to the meta-bolome level. As a consequence, explorative analysis (in contrast to purely hypothesis driven research procedures) has become applicable. However, numerous issues have to be considered for deriving meaningful results from Omics, and bioinformatics has to respect these in data analysis and interpretation. Aspects include sample type and quality, concise definition of the (clinical) question, and selection of samples ideally coming from thoroughly defined sample and data repositories. Omics suffers from a principal shortcoming, namely unbalanced sample-to-feature matrix denoted as "curse of dimensionality", where a feature refers to a specific gene or protein among the many thousands assayed in parallel in an Omics experiment. This setting makes the identification of relevant features with respect to a phenotype under analysis error prone from a statistical perspective. From this sample size calculation for screening studies and for verification of results from Omics, bioinformatics is essential. Here we present key elements to be considered for embedding Omics bioinformatics in a quality controlled workflow for Omics screening, feature identification, and validation. Relevant items include sample and clinical data management, minimum sample quality requirements, sample size estimates, and statistical procedures for computing the significance of findings from Omics bioinformatics in validation studies.*

7. Steyerberg EW, Schemper M, Harrell FE: Logistic regression modeling and the number of events per variable: selection bias dominates. *J Clin Epidemiol* (2011) 64(12): 1464-1465; author reply 1463-1464; https://doi.org/10.1016/j.jclinepi.2011.06.016

8. Waldhoer T, Heinzl H: Combining difference and equivalence test results in spatial maps. *Int J Health Geogr* (2011) 10(3): 3; https://doi.org/10.1186/1476-072X-10-3

   Abstract: *BACKGROUND: Regionally partitioned health indicator values are commonly presented in choropleth maps. Policymakers and health authorities use them among others for health reporting, demand planning and quality assessment. Quite often there are concerns whether the health situation in certain areas can be considered* *different or equivalent to a reference value. RESULTS: Highlighting statistically significant areas enables the statement that these areas differ from the reference value. However, this approach does not allow conclusions which areas are sufficiently close to the reference value, although these are crucial for health policy making as well. In order to overcome this weakness a combined integration of statistical difference and equivalence tests into choropleth maps is suggested and the approach is exemplified with health data of Austrian newborns. CONCLUSIONS: The suggested method will improve the interpretability of choropleth maps for policymakers and health authorities.*

2010

1. Dunkler D, Schemper M, Heinze G: Gene selection in microarray survival studies under possibly non-proportional hazards. *Bioinformatics* (2010) 26(6): 784-790; https://doi.org/10.1093/bioinformatics/btq035

   Abstract: *MOTIVATION: Univariate Cox regression (COX) is often used to select genes possibly linked to survival. With non-proportional hazards (NPH), COX could lead to under- or over-estimation of effects. The effect size measure c=P(T(1)<T(0)), i.e. the probability that a person randomly chosen from group G(1) dies earlier than a person from G(0), is independent of the proportional hazards (PH) assumption. Here we consider its generalization to continuous data c' and investigate the suitability of c' for gene selection. RESULTS: Under PH, c' is most efficiently estimated by COX. Under NPH, c' can be obtained by weighted Cox regression (WHE) or a novel method, concordance regression (CON). The least biased and most stable estimates were obtained by CON. We propose to use c' as summary measure of effect size to rank genes irrespective of different types of NPH and censoring patterns. AVAILABILITY: WHE and CON are available as R packages. CONTACT: georg.heinze@meduniwien.ac.at SUPPLEMENTARY INFORMATION: Supplementary Data are available at Bioinformatics online.*

2. Hainfellner JA, Heinzl H: Neuropathological biomarker candidates in brain tumors: key issues for translational efficiency. *Clin Neuropathol* (2010) 29(1): 41-54; https://doi.org/10.5414/npp29041

   Abstract: *Brain tumors comprise a large spectrum of rare malignancies in children and adults that are often associated with severe neurological symptoms and fatal outcome. Neuropathological tumor typing provides both prognostic and predictive tissue information which is the basis for optimal postoperative patient management and therapy. Molecular biomarkers may extend and refine prognostic and predictive information in a brain tumor case, providing more individualized and optimized treatment options. In the recent past a few neuropathological brain tumor biomarkers have translated smoothly into clinical use whereas many candidates show protracted translation. We investigated the causes of protracted translation of candidate brain tumor biomarkers. Considering the research environment from personal, social and systemic perspectives we identified eight determinants of translational success: methodology, funding, statistics, organization, phases of research, cooperation, self-reflection, and scientific progeny. Smoothly translating biomarkers are associated with low degrees of translational complexity whereas biomarkers with protracted translation are associated with high degrees. Key issues for translational efficiency of neuropathological brain tumor biomarker research seem to be related to (i) the strict orientation to the mission of medical research, that is the improval of medical practice as primordial purpose of research, (ii) definition of research priorities according to clinical needs, and (iii) absorption of translational complexities by means of operatively beneficial standards. To this end, concrete actions should comprise adequate scientific education of young investigators, and shaping of integrative diagnostics and therapy research both on the local level and the level of influential international brain tumor research platforms.*

3. Heinze G, Puhr R: Bias-reduced and separation-proof conditional logistic regression with small or sparse data sets. *Stat Med* (2010) 29(7-8): 770-777; https://doi.org/10.1002/sim.3794

   Abstract: *Conditional logistic regression is used for the analysis of binary outcomes when subjects are stratified into several subsets, e.g. matched pairs or blocks. Log odds ratio estimates are usually found by maximizing the conditional likelihood. This approach eliminates all strata-specific parameters by conditioning on the number of events within each stratum. However, in the analyses of both an animal experiment and a lung cancer case-control study, conditional maximum likelihood (CML) resulted in infinite odds ratio estimates and monotone likelihood. Estimation can be improved by using Cytel Inc.'s well-known LogXact software, which provides a median unbiased estimate and exact or mid-p confidence intervals. Here, we suggest and outline point and interval estimation based on maximization of a penalized conditional likelihood in the spirit of Firth's (Biometrika 1993; 80:27-38) bias correction method (CFL). We present comparative analyses of both studies, demonstrating some advantages of CFL over competitors. We report on a small-sample simulation study where CFL log odds ratio estimates were almost unbiased, whereas LogXact estimates showed some bias and CML estimates exhibited serious bias. Confidence intervals and tests based on the penalized conditional likelihood had close-to-nominal coverage rates and yielded highest power among all methods compared, respectively. Therefore, we propose CFL as an attractive solution to*

the stratified analysis of binary data, irrespective of the occurrence of monotone likelihood. A SAS program implementing CFL is available at: http://www.muw.ac.at/msi/biometrie/programs.

4. Karch R, Neumann F, Ullrich R, Heinze G, Neumuller J, Podesser BK, Neumann M: Methods from the theory of random heterogeneous media for quantifying myocardial morphology in normal and dilated hearts. *Ann Biomed Eng* (2010) 38(2): 308-318; https://doi.org/10.1007/s10439-009-9848-1

Abstract: *In the present study, descriptors from the theory of random heterogeneous media were used to characterize the morphology of the myocardial interstitial space in histological sections from hearts of healthy subjects and of patients with idiopathic dilated cardiomyopathy (DCM). Histological sections from resected DCM hearts (n = 9) were compared with donor hearts showing no signs of cardiac disease (n = 6). From control to DCM, the area fraction phi(1) of the interstitial space increased from 0.13 +/- 0.05 to 0.27 +/- 0.08, the chord-length z from 1.67 +/- 0.61 to 5.56 +/- 1.78 microm, the pore-size delta from 0.72 +/- 0.13 to 1.73 +/- 0.40 microm, the distance r (min) of the first local minimum in the two-point correlation function from 10.99 +/- 1.09 to 18.57 +/- 4.36 mum, whereas specific interface length s and decay-rate gamma of the lineal-path function decreased from 0.20 +/- 0.07 to 0.16 +/- 0.04 microm(-1) and from 0.39 +/- 0.09 to 0.16 +/- 0.05 microm(-1), respectively. All descriptors (except for s) were significantly different (p < 0.05) between control and DCM, reflecting an increasingly heterogeneous morphology in DCM hearts. Our results suggest that (1) descriptors originally developed to characterize the morphology of random heterogeneous media are well suited for histomorphometry of DCM, and (2) among the descriptors studied, either pore-size delta or chord-length z qualify best to discriminate between control and DCM hearts.*

5. Leffondre K, Wynant W, Cao Z, Abrahamowicz M, Heinze G, Siemiatycki J: A weighted Cox model for modelling time-dependent exposures in the analysis of case-control studies. *Stat Med* (2010) 29(7-8): 839-850; https://doi.org/10.1002/sim.3764

Abstract: *Many exposures investigated in epidemiological case-control studies may vary over time. The effects of these exposures are usually estimated using logistic regression, which does not directly account for changes in covariate values over time within individuals. By contrast, the Cox model with time-dependent covariates directly accounts for these changes over time. However, the over-sampling of cases in case-control studies, relative to controls, requires manipulating the risk sets in the Cox partial likelihood. A previous study showed that simple inclusion or exclusion of future cases in each risk set induces an under- or over-estimation bias in the regression parameters, respectively. We investigate the performance of a weighted Cox model that weights subjects according to age-conditional probabilities of developing the disease of interest in the source population. In a simulation study, the lifetime experience of a source population is first generated and a case-control study is then simulated within each population. Different characteristics of exposure are generated, including time-varying intensity. The results show that the estimates from the weighted Cox model are much less biased than the Cox models that simply include or exclude future cases, and are superior to logistic regression estimates in terms of bias and mean-squared error. An application to frequency-matched population-based case-control data on lung cancer illustrates similar differences in the estimated effects of different smoking variables. The investigated weighted Cox model is a potential alternative method to analyse matched or unmatched population-based case-control studies with time-dependent exposures.*

2009

1. Edler L, Lee JW, Mittlbock M, Niland J, Victor N: Computational statistics within clinical research. *Computational Statistics & Data Analysis* (2009) 53(3): 583-585; https://doi.org/10.1016/j.csda.2008.10.001

2. Heinzl H: Kaplan-Meier-Kurven und die Hazard Ratio. *krebs:hilfe!* (2009) 2:X-XI;

3. Mittlböck M: Editorial, Einführung in das statistische Testen. *krebs:hilfe!* (2009) 2:VI-VIII;

4. Schemper M: Book Review: Multivariable Model-building: A Pragmatic Approach to Regression Analysis based on Fractional Polynomials for Modelling Continous Variables. *Statistics in Medicine* (2009) 28(537-539;

5. Schemper M, Wakounig S, Heinze G: The estimation of average hazard ratios by weighted Cox regression. *Stat Med* (2009) 28(19): 2473-2489; https://doi.org/10.1002/sim.3623

   Abstract: *Often the effect of at least one of the prognostic factors in a Cox regression model changes over time, which violates the proportional hazards assumption of this model. As a consequence, the average hazard ratio for such a prognostic factor is under- or overestimated. While there are several methods to appropriately cope with non-proportional hazards, in particular by including parameters for time-dependent effects, weighted estimation in Cox regression is a parsimonious alternative without additional parameters. The methodology, which extends the weighted k-sample logrank tests of the Tarone-Ware scheme to models with multiple, binary and continuous covariates, has been introduced in the nineties of the last century and is further developed and re-evaluated in this contribution. The notion of an average hazard ratio is defined and its connection to the effect size measure P(X<Y) is emphasized. The suggested approach accomplishes estimation of intuitively interpretable average hazard ratios and provides tools for inference. A Monte Carlo study confirms the satisfactory performance. Advantages of the approach are exemplified by comparing standard and weighted analyses of an international lung cancer study. SAS and R programs facilitate application.*

2008

1. Heinze G, Dunkler D: Avoiding infinite estimates of time-dependent effects in small-sample survival studies. *Stat Med* (2008) 27(30): 6455-6469; https://doi.org/10.1002/sim.3418

   Abstract: *We address the phenomenon of monotone likelihood in Cox regression with time-dependent effects. Monotone likelihood occurs in the fitting process of a Cox model if at least one parameter estimate diverges to +/- infinity. We show that the probability of monotone likelihood is increased by the inclusion of time-dependent effects, particularly in small samples with several unbalanced and highly predictive covariates, and with a high percentage of censoring. Firth's bias reduction procedure was shown to provide an ideal solution to monotone likelihood. Here we extend his idea to Cox regression with time-dependent effects. By penalized maximum likelihood estimation, finite hazard ratio estimates of constant and time-dependent effects can be obtained. Penalized likelihood ratio tests and profile penalized likelihood confidence intervals are proposed as tools for inference. A Monte Carlo study of Cox regression with time-dependent effects confirms advantages of Firth-corrected (FC) over standard Cox analysis in terms of average bias and median absolute deviation. We also compare the FC and standard Cox approaches by means of analyses of two studies with time-dependent effects. An SAS macro and an R package for FC Cox regression with time-varying covariates and time-dependent effects are available at: http://www.muw.ac.at/msi/biometrie/programs.*

2. Heinzl H, Mittlbock M, Edler L: Technical uncertainty in the back-calculation of occupational exposure to dioxins. *Stat Med* (2008) 27(12): 2214-2233; https://doi.org/10.1002/sim.3074

   Abstract: *Members of a cohort of workers in chemical industry (the so-called Boehringer cohort) exposed to 2, 3, 7, 8-tetrachlorodibenzo-para-dioxin (TCDD) from 1950 to 1984 were subject in the years 1985-1986 and 1992-1994 to an extensive biomonitoring programme on the TCDD levels of the individual workers. For establishing a dose-response relationship between TCDD-exposure and potentially carcinogenic response, the individual TCDD concentration-time courses had to be back-calculated over a period of up to more than four decades. Two back-calculations were attempted for this sophisticated modelling and estimation task, both based on the same toxicokinetic model but yielding different results. We demonstrate here by means of a computer simulation study that these differences could be plausibly explained by the so-called technical uncertainty caused by the employment of differently statistical estimation techniques. We show that the estimation techniques perform particularly differently in the presence of workplace misclassification and TCDD measurement error, two complications of exposure assessment that are with high probability affecting concurrently that cohort's data. We conclude that technical uncertainty sensibly enlarges the pool of possible explanations for contradictory empirical results of complex modelling and estimation approaches and should be considered as an obligatory uncertainty analysis step after the primary risk analysis evaluation in epidemiological and environmental studies.*

3. Mittlbock M: Critical Appraisal of Randomized Clinical Trials: Can We Have Faith in the Conclusions? *Breast Care (Basel)* (2008) 3(5): 341-346; https://doi.org/10.1159/000157168

   Abstract: *Randomized clinical trials (RCTs) are the most appropriate research design for studying the effectiveness of a specific intervention. Its results are considered as the highest 'level of evidence'. Published reports on RCTs have already succeeded in a peer review process, but still there can be undetected major deficiencies of the study that may question the reported outcome. It is still up to the readers to assess the quality of publications and to question if the published results apply to their patients. The major points of such a critical appraisal process are reviewed and discussed with a focus on breast cancer studies.*

2007

1. Dunkler D, Michiels S, Schemper M: Gene expression profiling: does it add predictive accuracy to clinical characteristics in cancer prognosis? *Eur J Cancer* (2007) 43(4): 745-751; https://doi.org/10.1016/j.ejca.2006.11.018

   Abstract: *It is widely accepted that gene expression classifiers need to be externally validated by showing that they predict the outcome well enough on other patients than those from whose data the classifier was derived. Unfortunately, the gain in predictive accuracy by the classifier as compared to established clinical prognostic factors often is not quantified. Our objective is to illustrate the application of appropriate statistical measures for this purpose. In order to compare the predictive accuracies of a model based on the clinical factors only and of a model based on the clinical factors plus the gene classifier, we compute the decrease in predictive inaccuracy and the proportion of explained variation. These measures have been obtained for three studies of published gene classifiers: for survival of lymphoma patients, for survival of breast cancer patients and for the diagnosis of lymph node metastases in head and neck cancer. For the three studies our results indicate varying and possibly small added explained variation and predictive accuracy due to gene classifiers. Therefore, the gain of future gene classifiers should routinely be demonstrated by appropriate statistical measures, such as the ones we recommend.*

2. Heinzl H, Benner A, Ittrich C, Mittlbock M: Proposals for sample size calculation programs. *Methods Inf Med* (2007) 46(6): 655-661; https://doi.org/10.3414/me0295

   Abstract: *OBJECTIVES: Numerous sample size calculation programs are available nowadays. They include both commercial products as well as public domain and open source applications. We propose modifications for these programs in order to even better support statistical consultation during the planning stage of a two-armed clinical trial. METHODS: Directional two-sided tests are commonly used for two-armed clinical trials. This may lead to a non-negligible Type III error risk in a severely underpowered study. In the case of a reasonably sized study the question for the so-called auxiliary alternative may evolve. RESULTS: We propose that sample size calculation programs should be able to compute i) Type III errors and the so-called q-values, ii) minimum sample sizes required to keep the q-values below pre-specified levels, and iii) detectable effect sizes of the so-called auxiliary alternatives. CONCLUSIONS: Proposals i and ii are intended to help prevent irresponsibly underpowered clinical trials, whereas the proposal iii is meant as additional assistance for the planning of reasonably sized clinical trials.*

3. Heinzl H, Mittlböck M: Eine Einführung zu chirurgischen Lernkurvenstudien - Aus Sicht der medizinischen Statistik und Dokumentation. *Forum der Medizin_Dokumentation und Medizin_Informatik (mdi)* (2007) 9(1): 30-33;

4. Heinzl H, Mittlbock M, Edler L: On the translation of uncertainty from toxicokinetic to toxicodynamic models - The TCDD example. *Chemosphere* (2007) 67(9): S365-S374; https://doi.org/10.1016/j.chemosphere.2006.05.130

   Abstract: *When estimating human health risks from exposure to TCDD using toxicokinetic and toxicodynarnic models, it is important to understand how model choice and assumptions necessary for modeling add to the uncertainty of risk estimates. Several toxicokinetic models have been proposed for the risk assessment of dioxins, in particular the elimination kinetics in humans has been a matter of constant debate. For a long time, a simple linear elimination kinetics has been common choice. Thus, it was used for the statistical analysis of the largest occupationally exposed cohort, the German Boehringer cohort.*
   We challenge this assumption by considering, amongst others, a nonlinear modified Michaelis-Menten-type elimination kinetics, the so-called Carrier kinetics. Using the area under the lipid TCDD concentration time curve as dose metrics, we model the time to cancer-related death using the Cox proportional hazards model as toxicodynamic model. This risk assessment set-up was simulated in order to quantify uncertainty of both the dose (TCDD body burden) and the risk estimates, depending on the use of the kinetic model, variations of carcinogenic effect of TCDD and variations of latency period (lag time). If past exposure is estimated assuming a linear elimination kinetics although a Carrier kinetics actually holds, then high exposures in reality will be underestimated through statistical analysis and low exposures will be overestimated, respectively. This bias will

carry over on the estimated individual concentration-time curves and the therefrom derived TCDD dose metric values. Using biased dose values when estimating a dose-response relationship will finally lead to biased risk estimates. The extent of bias and the decrease of precision are quantified in selected scenarios through this simulation approach. Our findings are in concordance with recent results in the field of dioxin risk assessment. They also reinforce the general demand for the scheduled uncertainty assessments in risk analyses. (c) 2006 Elsevier Ltd. All rights reserved.

5. Lehr S, Schemper M: Parsimonious analysis of time-dependent effects in the Cox model. *Stat Med* (2007) 26(13): 2686-2698; https://doi.org/10.1002/sim.2742

Abstract: *Cox's proportional hazards model can be extended to accommodate time-dependent effects of prognostic factors. We briefly review these extensions along with their varying degrees of freedom. Spending more degrees of freedom with conventional procedures (a priori defined interactions with simple functions of time, restricted natural splines, piecewise estimation for partitions of the time axis) allows the fitting of almost any shape of time dependence but at an increased risk of over-fit. This results in increased width of confidence intervals of time-dependent hazard ratios and in reduced power to confirm any time-dependent effect or even any effect of a prognostic factor. By means of comparative empirical studies the consequences of over-fitting time-dependent effects have been explored. We conclude that fractional polynomials, and similarly penalized likelihood approaches, today are the methods of choice, avoiding over-fit by parsimonious use of degrees of freedom but also permitting flexible modelling if time dependence of a usually a priori unknown shape is present in a data set. The paradigm of a parsimonious analysis of time-dependent effects is exemplified by means of a gastric cancer study.*

2006

1. Draxler W, Mittlböck M: Basic principles in the planning of clinical trials in surgical oncology. *European Surgery* (2006) 38(1): 27-32; https://doi.org/10.1007/s10353-006-0211-6

   Abstract: *Background: ICH (International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use) provides guidelines on the implementation of clinical trials. All study participants are obliged to follow these guidelines in line with "Good Clinical Practice".*
   Methods: The main features of a clinical study include the following items: Background and general aims, specific objectives, patient selection criteria, treatment schedules, methods of patient evaluation, trial design, registration and randomization of patients, patient consent, required size of study, monitoring of trial progress, forms and data handling, protocol deviations, plans for statistical analysis and administrative responsibilities.
   Results: All items mentioned above should already be discussed in the planning stage of a clinical trial and addressed in the study protocol. The study protocol provides a guideline for any person involved in the trial.
   Conclusions: For the success of a clinical trial, it is especially important to have a clear and exact definition of the study hypotheses and to choose primary and secondary endpoints very carefully.

2. Heinze G: A comparative investigation of methods for logistic regression with separated or nearly separated data. *Stat Med* (2006) 25(24): 4216-4226; https://doi.org/10.1002/sim.2687

   Abstract: *In logistic regression analysis of small or sparse data sets, results obtained by classical maximum likelihood methods cannot be generally trusted. In such analyses it may even happen that the likelihood meets the convergence criteria while at least one parameter estimate diverges to +/-infinity. This situation has been termed 'separation', and it typically occurs whenever no events are observed in one of the two groups defined by a dichotomous covariate. More generally, separation is caused by a linear combination of continuous or dichotomous covariates that perfectly separates events from non-events. Separation implies infinite or zero maximum likelihood estimates of odds ratios, which are usually considered unrealistic. I provide some examples of separation and near-separation in clinical data sets and discuss some options to analyse such data, including exact logistic regression analysis and a penalized likelihood approach. Both methods supply finite point estimates in case of separation. Profile penalized likelihood confidence intervals for parameters show excellent behaviour in terms of coverage probability and provide higher power than exact confidence intervals. General advantages of the penalized likelihood approach are discussed.*

3. Heinze G, Schemper M: A permutation test for inference in logistic regression with small- and moderate-sized data sets. *Stat Med* (2006) 25(4): 719; https://doi.org/10.1002/sim.2281

4. Mittlböck M, Heinzl H: A simulation study comparing properties of heterogeneity measures in meta-analyses. *Stat Med* (2006) 25(24): 4321-4333; https://doi.org/10.1002/sim.2692

   Abstract: *The assessment of heterogeneity or between-study variance is an important issue in meta-analysis. It determines the statistical methods to be used and the interpretation of the results. Tests of heterogeneity may be misleading either due to low power for sparse data or to the detection of irrelevant amounts of heterogeneity when many studies are involved. In the former case, notable heterogeneity may remain unconsidered and an unsuitable model may be chosen and the latter case may lead to unnecessary complex analyses strategies. Measures of heterogeneity are better suited to determine appropriate analyses strategies. We review two measures with different scaling and compare them with the heterogeneity test. Estimates of the within-study variance are discussed and a new total information measure is introduced. Various properties of the quantities in question are assessed by a simulation study. Heterogeneity test and measures are not directly related to the amount of between-study variance but to the relative increase of variance due to heterogeneity. It is more favourable to base the within-study variance estimate on the squared weights of individual studies than on the sum of weights. A heterogeneity measure scaled to a fixed interval needs reference values for proper interpretation. A measure defined by the relation of between- to within-study variance has a more natural interpretation but no upper limit. Both measures are quantifications of the impact of heterogeneity on the meta-analysis result as both depend on the variance of the individual study effects and thus on the number of patients in the studies.*

2005

1. Heinzl H, Mittlböck M, Edler L: On the role of ex post uncertainty assessment for risk management. *International Journal of Risk Assessment and Management* (2005) 5(2,3,4): 206-215;

   Abstract: *A risk management decision whether a chemical compound present in the environment is potentially hazardous to human health will be among others based on large-scaled statistical analyses of empirical data of various sources. Observations from highly exposed and well-documented occupational cohorts are usually a matter of particular interest. Due to scientific progress, various new aspects will inevitably emerge in the course of time. These new aspects are not always in full accordance with the assumptions of the original statistical analyses so that concern may evolve whether the original foundation of the risk assessment is still valid in the main. We suggest to account for such doubts by an ex post uncertainty assessment of the original statistical analyses. The case of the risk assessment whether 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) is a human carcinogen is used as example. We call for the scheduling of future ex post uncertainty assessments already at the time of the risk management decision.*

2. Heinzl H, Waldhor T, Mittlbock M: Careful use of pseudo R-squared measures in epidemiological studies. *Stat Med* (2005) 24(18): 2867-2872; https://doi.org/10.1002/sim.2168

   Abstract: *Many epidemiological research problems deal with large numbers of exposed subjects of whom only a small number actually suffers the adverse event of interest. Such rare events data can be analysed by employing an approximate Poisson model. The objective of this study is to challenge the interpretability of the corresponding Poisson pseudo R-squared measure. It will lack sensible interpretation whenever the approximate Poisson outcome is generated by counting the number of events within covariate patterns formed by cross-tabulating categorical covariates. The failure is caused by the immanent arbitrariness in the definition of the covariate patterns, that is, independent Bernoulli events, B(1,pi), are arbitrarily combined into binomially distributed ones, B(n,pi), which are then approximated by the Poisson model.*

2004

1. Edler L, Heinzl H, Mittlböck M: Carrying-over Toxicokinetic Model Uncertainty into Cancer Risk Estimates: The TCDD Example. *Organohalogen Compounds* (2004) 66:3356-3362;

2. Mittlböck M: Book Review zu Klein,JP., Moeschberger,ML. (2003). Survival Analysis: Techniques for Censored an Truncated Data. Springer, New York. *Biometrical Journal* (2004) 46(3): 379;

2003

1. Edler L, Heinzl H: Toxicokinetic modeling for environmental health problems. *Environmetrics* (2003) 14(2): 193-202; https://doi.org/10.1002/env.576

   Abstract: *Toxicokinetic (TK) and physiologically based toxicokinetic (PB-TK) models enable tissue dosimetry and the definition of the target organ dose after exposure to exogenous toxic compounds. This has qualified PB-TK models for extrapolation from the experimental animal to the human, from high to low doses, between routes of exposure, between patterns of exposure, and also between robust and susceptible sub-populations. We show how PB-TK models are constructed, what assumptions are needed, which mathematical methods are used for model building and by which statistical methods one may assess both model fit and uncertainty of the modeling itself. We will address in particular a new generation of PB-TK models which include age-dependent model parameters to model life-long human exposure with an example from the exposure to dioxins. Finally, we define the role of PB-TK models for the identification of human health effects after exposure to chemicals in risk assessment. Copyright (C) 2003 John Wiley Sons, Ltd.*

2. Heinze G, Gnant M, Schemper M: Exact log-rank tests for unequal follow-up. *Biometrics* (2003) 59(4): 1151-1157; https://doi.org/10.1111/j.0006-341x.2003.00132.x

   Abstract: *The asymptotic log-rank and generalized Wilcoxon tests are the standard procedures for comparing samples of possibly censored survival times. For comparison of samples of very different sizes, an exact test is available that is based on a complete permutation of log-rank or Wilcoxon scores. While the asymptotic tests do not keep their nominal sizes if sample sizes differ substantially, the exact complete permutation test requires equal follow-up of the samples. Therefore, we have developed and present two new exact tests also suitable for unequal follow-up. The first of these is an exact analogue of the asymptotic log-rank test and conditions on observed risk sets, whereas the second approach permutes survival times while conditioning on the realized follow-up in each group. In an empirical study, we compare the new procedures with the asymptotic log-rank test, the exact complete permutation test, and an earlier proposed approach that equalizes the follow-up distributions using artificial censoring. Results confirm highly satisfactory performance of the exact procedure conditioning on realized follow-up, particularly in case of unequal follow-up. The advantage of this test over other options of analysis is finally exemplified in the analysis of a breast cancer study.*

3. Heinze G, Ploner M: Fixing the nonconvergence bug in logistic regression with SPLUS and SAS. *Comput Methods Programs Biomed* (2003) 71(2): 181-187; https://doi.org/10.1016/s0169-2607(02)00088-3

   Abstract: *When analyzing clinical data with binary outcomes, the parameter estimates and consequently the odds ratio estimates of a logistic model sometimes do not converge to finite values. This phenomenon is due to special conditions in a data set and known as 'separation'. Statistical software packages for logistic regression using the maximum likelihood method cannot appropriately deal with this problem. A new procedure to solve the problem has been proposed by Heinze and Schemper (Stat. Med. 21 (2002) pp. 2409-3419). It has been shown that unlike the standard maximum likelihood method, this method always leads to finite parameter estimates. We developed a SAS macro and an SPLUS library to make this method available from within one of these widely used statistical software packages. Our programs are also capable of performing interval estimation based on profile penalized log likelihood (PPL) and of plotting the PPL function as was suggested by Heinze and Schemper (Stat. Med. 21 (2002) pp. 2409-3419).*

4. Heinze G, Schemper M: Comparing the importance of prognostic factors in Cox and logistic regression using SAS. *Computer Methods and Programs in Biomedicine* (2003) 71(2): 155-163; https://doi.org/10.1016/S0169-2607(02)00077-9

   Abstract: *Two SAS macro programs are presented that evaluate the relative importance of prognostic factors in the proportional hazards regression model and in the logistic regression model. The importance of a prognostic factor is quantified by the proportion of variation in the outcome attributable to this factor. For proportional hazards regression, the program %RELIMPCR uses the recently proposed measure V to calculate the proportion of explained variation (PEV). For the logistic model, the R-2 measure based on squared raw residuals is used by the*

program %RELIMPLR. Both programs are able to compute marginal and partial PEV, to compare PEVs of factors, of groups of factors, and even to compare PEVs of different models. The programs use a bootstrap resampling scheme to test differences of the PEVs of different factors. Confidence limits for P-values are provided. The programs further allow to base the computation of PEV on models with shrinked or bias-corrected parameter estimates. The SAS macros are freely available at www.akh-wien.ac.at/imc/biometrie/relimp. (C) 2002 Elsevier Science Ireland Ltd. All rights reserved.

5. Heinzl H, Mittlbock M: Pseudo R-squared measures for Poisson regression models with over- or underdispersion. *Computational Statistics & Data Analysis* (2003) 44(1-2): 253-271; https://doi.org/10.1016/S0167-9473(03)00062-8

Abstract: *The Poisson regression model is frequently used to analyze count data. Pseudo R-squared measures for Poisson regression models have recently been proposed and bias adjustments recommended in the presence of small samples and/or a large number of covariates. In practice, however, data are often over- or sometimes even underdispersed as compared to the standard Poisson model. The definition of Poisson R-squared measures can be applied in these situations as well, albeit with bias adjustments accordingly adapted. These adjustments are motivated by arguments of quasi-likelihood theory. Properties of unadjusted and adjusted R-squared measures are studied by simulation under standard Poisson; over- and underdispersed Poisson regression models and their use is exemplified and discussed with popcorn data. (C) 2003 Elsevier B.V. All rights reserved.*

6. Nardi A, Schemper M: Comparing Cox and parametric models in clinical studies. *Stat Med* (2003) 22(23): 3597-3610; https://doi.org/10.1002/sim.1592

Abstract: *Parametric models are only occasionally used in the analysis of clinical studies of survival although they may offer advantages over Cox's model. In this paper, we report experiences that we have made fitting parametric models to data sets from different clinical trials mainly performed at the Vienna University Medical School. We emphasize the role of residuals for discriminating among candidate models and judging their goodness of fit. The effect of misspecification of the baseline distribution on parameter estimates and testing has been explored. The results from parametric analyses have always been contrasted with those from Cox's model.*

7. Schemper M: Predictive accuracy and explained variation. *Stat Med* (2003) 22(14): 2299-2308; https://doi.org/10.1002/sim.1486

Abstract: *Measures of the predictive accuracy of regression models quantify the extent to which covariates determine an individual outcome. Explained variation measures the relative gains in predictive accuracy when prediction based on covariates replaces unconditional prediction. A unified concept of predictive accuracy and explained variation based on the absolute prediction error is presented for models with continuous, binary, polytomous and survival outcomes. The measures are given both in a model-based formulation and in a formulation directly contrasting observed and expected outcomes. Various aspects of application are demonstrated by examples from three forms of regression models. It is emphasized that the likely degree of absolute or relative predictive accuracy often is low even if there are highly significant and relatively strong covariates.*

8. Schemper M: Some Recent Developments In Survival Analysis. *Craiova Medicala Journal* (2003) 5(3): 22-27;

9. Schreiner W, Karch R, Neumann M, Neumann F, Roedler SM, Heinze G: Heterogeneous perfusion is a consequence of uniform shear stress in optimized arterial tree models. *J Theor Biol* (2003) 220(3): 285-301; https://doi.org/10.1006/jtbi.2003.3136

Abstract: *Using optimized computer models of arterial trees we demonstrate that flow heterogeneity is a necessary consequence of a uniform shear stress distribution. Model trees are generated and optimized under different modes of boundary conditions. In one mode flow is delivered to the tissue as homogeneously as possible. Although this primary goal can be achieved, resulting shear stresses between blood and the vessel walls show very large spread. In a second mode, models are optimized under the condition of uniform shear stress in all segments which in turn renders flow distribution heterogeneous. Both homogeneous perfusion and uniform shear stress are desirable goals in real arterial trees but each of these goals can only be approached at the expense of the other.*

While the present paper refers only to optimized models, we assume that this dual relation between the heterogeneities in flow and shear stress may represent a more general principle of vascular systems.

2002

1. Heinze G, Ploner M: SAS and SPLUS programs to perform Cox regression without convergence problems. *Computer Methods and Programs in Biomedicine* (2002) 67(3): 217-223; https://doi.org/10.1016/S0169-2607(01)00149-3

   Abstract: *When analyzing survival data, the parameter estimates and consequently the relative risk estimates of a Cox model sometimes do not converge to finite values. This phenomenon is due to special conditions in a data set and is known as 'monotone likelihood'. Statistical software packages for Cox regression using the maximum likelihood method cannot appropriately deal with this problem. A new procedure to solve the problem has been proposed by G. Heinze, M. Schemper, A solution to the problem of monotone likelihood in Cox regression, Biometrics :57 (2001). It has been shown that unlike the standard maximum likelihood method, this method always leads to finite parameter estimates. We developed a SAS macro and an SPLUS library to make this method available from within one of these widely used statistical software packages. Our programs are also capable of performing interval estimation based on profile penalized log likelihood (PPL) and of plotting the PPL function as was suggested by G. Heinze, M. Schemper, A solution to the problem of monotone likelihood in Cox regression, Biometrics 57 (2001). (C) 2002 Elsevier Science Ireland Ltd. All rights reserved.*

2. Heinze G, Schemper M: A solution to the problem of separation in logistic regression. *Stat Med* (2002) 21(16): 2409-2419; https://doi.org/10.1002/sim.1047

   Abstract: *The phenomenon of separation or monotone likelihood is observed in the fitting process of a logistic model if the likelihood converges while at least one parameter estimate diverges to +/- infinity. Separation primarily occurs in small samples with several unbalanced and highly predictive risk factors. A procedure by Firth originally developed to reduce the bias of maximum likelihood estimates is shown to provide an ideal solution to separation. It produces finite parameter estimates by means of penalized maximum likelihood estimation. Corresponding Wald tests and confidence intervals are available but it is shown that penalized likelihood ratio tests and profile penalized likelihood confidence intervals are often preferable. The clear advantage of the procedure over previous options of analysis is impressively demonstrated by the statistical analysis of two cancer studies.*

3. Heinzl H, Edler L: Assessing Uncertainty in a Toxicokinetic Model for Human Lifetime Exposure to TCDD. *Organohalogen Compounds* (2002) 59:355-358;

4. Heinzl H, Mittlböck M: Adjusted R2 Measures for the Inverse Gaussian Regression Model. *Computational Statistics* (2002) 17(525-544;

   Abstract: *The R² measure is a commonly used tool for assessing the predictive ability of a linear regression model. It quantifies the amount of variation in the outcome variable, which is explained by the covariates. Various attempts have been made to carry the R² definition to other types of regression models as well. Here, two different R² measure definitions for the Inverse Gaussian regression model will be studied. They are motivated by deviance and sums-of-squares residuals. Depending on sample size and number of covariates fitted, these R² measures may show substantially inflated values, and a proper bias-adjustment is necessary. Several possible adjusted R² measure definitions for the Inverse Gaussian regression model will be compared in a simulation study. The use of adjusted R² measures is recommended in general.*

5. Heinzl H, Stare J, Mittlbock M: A measure of dependence for the stratified Cox proportional hazards regression model. *Biometrical Journal* (2002) 44(6): 671-682; https://doi.org/10.1002/1521-4036(200209)44:6<671::AID-BIMJ671>3.0.CO;2-E

   Abstract: *KENT and O'QUIGLEY (1988) apply the concept of information gain to measure both global and partial dependence between explanatory variables and a censored response within the framework of the proportional hazards regression model of Cox (1972). The definition of this measure is extended to cover also the stratified Cox model.*

6. Mittlböck M: Calculating adjusted R(2) measures for Poisson regression models. *Comput Methods Programs Biomed* (2002) 68(3): 205-214; https://doi.org/10.1016/s0169-2607(01)00173-0

Abstract: *In regression models not only the parameter estimates and significances of explanatory variables are of interest, but also the degree to which variation in the dependent variable can be explained by covariates. In recent publications, an R(2) measure based on deviance was recommended for Poisson regression models, one of the most frequently used modelling tools in epidemiological studies. However, when sample size is small relative to the number of covariates in the model, simple R(2) measures may be seriously inflated and may need to be adjusted according to the number of covariates in the model. We present a SAS-macro that calculates adjustments for the R(2) measures in Poisson regression models based on log-likelihood and on sums of squares. The proposed measures are applied to real data sets and their performance is discussed.*

7. Mittlböck M, Heinzl H: Measures of explained variation in gamma regression models. *COMMUNICATIONS IN STATISTICS-SIMULATION AND COMPUTATION* (2002) 31(1): 61-73;

Abstract: *The common R-2 measure provides a useful means to quantify the degree to which variation in the dependent variable can be explained by the covariates in a linear regression model. Recently, there have been various attempts to apply the definition of the R-2 measure to generalized linear models. This paper studies two different R-2 measure definitions for the gamma regression model. These measures are related to deviance and sum-of-squares residuals. Depending on the sample size and the number of covariates fitted, so-called unadjusted R-2 measures may be substantially inflated, and the use of adjusted R-2 measures is then preferred. We study several known adjustments previously proposed for R-2 measures in regression models and illustrate the effect on the two unadjusted R-2 measures for the gamma regression model. Comparing the resulting measures with underlying population values, we find the best adjustment via simulation.*

8. Mittlbock M, Schemper M: Explained variation for logistic regression - Small sample adjustments, confidence intervals and predictive precision. *Biometrical Journal* (2002) 44(3): 263-272; https://doi.org/10.1002/1521-4036(200204)44:3<263::AID-BIMJ263>3.0.CO;2-7

Abstract: *The proportion of explained variation in logistic regression can be expressed by the multiple R 2 originally developed for the general linear model (cf. MITTLBOCK and SCHEMPER (1996)). In this paper we present a detailed investigation of this measure in small samples and/or with many covariates and propose either of two adjustments, one being a direct analogue of R-adj(2) of the general linear model, and the other being based on shrinkage. Furthermore, we explore the use of bootstrap confidence intervals and give a table of the expected variability of estimates of explained variation for samples of varying sizes. We recommend to quantify gains of predictive precision due to prognostic factors by both relative and absolute measures. For binary outcomes the components of the relative measure, R-2, are suitable absolute measures of predictive precision. They are interpretable as average absolute residuals conditional on using prognostic factors and without such information. We motivate application of the presented measures by the statistical analysis of a study of physical characteristics of urine possibly related to the presence of calcium oxalate crystals.*

2001

1. Gall W, Heinzl H, Sachs P: Extracting a statistical data matrix from electronic patient records. *Comput Methods Programs Biomed* (2001) 66(2-3): 153-166; https://doi.org/10.1016/s0169-2607(00)00130-9

   Abstract: *This paper describes the processing and transformation of medical data from a clinical database to a statistical data matrix. Precise extraction and linking tools must be available for the desired data to be processed for statistical purposes. We show that flexible mechanisms are required for the different types of users, such as physicians and statisticians. In our retrieval tools we use logical queries based on operands and operators. The paper describes the method and appliance of the operators with which the desired matrix is created through a process of selection and linking. Examples with a Kaplan-Meier function and time-dependent covariables demonstrate how our model is useful for different user groups.*

2. Heinze G, Schemper M: A solution to the problem of monotone likelihood in Cox regression. *Biometrics* (2001) 57(1): 114-119; https://doi.org/10.1111/j.0006-341x.2001.00114.x

   Abstract: *The phenomenon of monotone likelihood is observed in the fitting process of a Cox model if the likelihood converges to a finite value while at least one parameter estimate diverges to +/- infinity. Monotone likelihood primarily occurs in small samples with substantial censoring of survival times and several highly predictive covariates. Previous options to deal with monotone likelihood have been unsatisfactory. The solution we suggest is an adaptation of a procedure by Firth (1993, Biometrika 80, 27-38) originally developed to reduce the bias of maximum likelihood estimates. This procedure produces finite parameter estimates by means of penalized maximum likelihood estimation. Corresponding Wald-type tests and confidence intervals are available, but it is shown that penalized likelihood ratio tests and profile penalized likelihood confidence intervals are often preferable. An empirical study of the suggested procedures confirms satisfactory performance of both estimation and inference. The advantage of the procedure over previous options of analysis is finally exemplified in the analysis of a breast cancer study.*

3. Heinzl H, Tempfer C: A cautionary note on segmenting a cyclical covariate by minimum P-value search. *Computational Statistics & Data Analysis* (2001) 35(4): 451-461; https://doi.org/10.1016/S0167-9473(00)00023-2

   Abstract: *Recently, menstrual status at the time of surgery was suggested to be a potential prognostic factor for survival in premenopausal women suffering from breast cancer. That is, surgery should be avoided in a certain segment of the menstrual cycle. However, besides that different authors claimed different segments to be hazardous, the alleged influence on survival could hardly be confirmed by subsequent studies. Statistical arguments could provide an explanation for these contradictory findings. Splitting a cyclical covariate into two segments is analogous to dichotomizing a continuous covariate. Given that this splitting is based on a minimum P-value search, the actual type I error rate will be much higher than the nominal one due to multiple testing. A simulation study has been performed to gain insight into the problem. (C) 2001 Elsevier Science B.V. All rights reserved.*

4. Mittlböck M, Heinzl H: A note on R-2 measures for Poisson and logistic regression models when both models are applicable. *Journal of Clinical Epidemiology* (2001) 54(1): 99-103; https://doi.org/10.1016/S0895-4356(00)00292-4

   Abstract: *The aim of many epidemiological studies is the regression of a dichotomous outcome (e.g., death or affection by a certain disease) on prognostic covariables. Thereby the Poisson regression model is often used alternatively to the logistic regression model. Modelling the number of events and individual outcomes, respectively, both models lead to nearly the same results concerning the parameter estimates and their significances. However. when calculating the proportion of explained variation, quantified by an R-2 measure, a large difference between both models usually occurs. We illustrate this difference by an example and explain it with theoretical arguments. We conclude, the R-2 measure of the Poisson regression quantifies the predictability of event rates, but it is not adequate to quantify the predictability of the outcome of individual observations. (C) 2001 Elsevier Science Inc. All rights reserved.*

5. Stare J, Harrell FE, Heinzl H: BJ: an S-Plus program to fit linear regression models to censored data using the Buckley-James method. *Computer Methods and Programs in Biomedicine* (2001) 64(1): 45-52; https://doi.org/10.1016/S0169-2607(00)00083-3

   Abstract: *Most researchers are Familiar with ordinary multiple regression models, most commonly fitted using the method of least squares. The method of Buckley and James (J. Buckley, I. James, Linear regression with censored data, Biometrika 66 (1979) 429-436.) is an extension of least squares for fitting multiple regression models when the response variable is right-censored as in the analysis of survival time data. The Buckley-James method has been shown to have good statistical properties under usual regularity conditions (T.L. Lai, Z. Ying, Large sample theory of a modified Buckley-James estimator for regression analysis with censored data, Ann. Stat. 19 (1991) 1370-1402.). Nevertheless, even after 20 years of its existence, it is almost never used in practice. We believe that this is mainly due to lack of software and we describe here an S-Plus program that through its inclusion in a public domain function library fully exploits the power of the S-Plus programming environment. This environment provides multiple facilities for model specification, diagnostics, statistical inference, and graphical depiction of the model lit. (C) 2001 Elsevier Science Ireland Ltd. All rights reserved.*

6. Waldhör T, Mittlböck M, Haidinger G, Zidek T, Schober E: Partial R 2 -Values Based On Deviance Residuals In Poisson Regression Models. *Information, Biometrie und Epidemiologie in Medizin und Biologie* (2001) 32(4): 341-347;

2000

1. Heinzl H: Using SAS to calculate the Kent and O'Quigley measure of dependence for Cox proportional hazards regression model. *Computer Methods and Programs in Biomedicine* (2000) 63(1): 71-76; https://doi.org/10.1016/S0169-2607(00)00073-0

   Abstract: *Kent and O'Quigley (1988) apply the concept of information gain to define a measure of dependence (R-squared measure) between explanatory variables and a censored response variable within the framework of the Cox model, Two SAS macros to calculate this measure are presented. The first one is based on a Newton-Raphson search and makes use of the SAS IML procedure. The second one is a simple grid search using SAS DATA steps and Base-SAS procedures. (C) 2000 Elsevier Science Ireland Ltd. All rights reserved.*

2. Heinzl H: Dangers of Using 'Optimal' Cutpoints in the Evaluation of Cyclical Prognostic Factors. New Approaches in Applied Statistics. *New Approaches in Applied Statistics* (2000) 16:135-143;

3. Mittlböck M, Waldhör T: Adjustments for R-2-measures for Poisson regression models. *Computational Statistics & Data Analysis* (2000) 34(4): 461-472; https://doi.org/10.1016/S0167-9473(99)00113-9

   Abstract: *In regression models not only the parameter estimates and significances of explanatory variables are of interest, but also the degree to which variation in the dependent variable can be explained by covariates. In recent publications an R-2-measure based on deviance was recommended for Poisson regression models, one of the most frequently used modelling tools in epidemiological studies. However, when sample size is small relative to the number of covariates in the model, simple R-2-measures may be seriously inflated and may need to be adjusted according to the number of covariates in the model. Two new adjustments for the R-2-measure in Poisson regression models based on deviance residuals are presented and compared by simulation with population values. The proposed measures are also applied to real data sets. (C) 2000 Elsevier Science B.V. All rights reserved.*

4. Schemper M, Henderson R: Predictive accuracy and explained variation in Cox regression. *Biometrics* (2000) 56(1): 249-255; https://doi.org/10.1111/j.0006-341x.2000.00249.x

   Abstract: *We suggest a new measure of the proportion of the variation of possibly censored survival times explained by a given proportional hazards model. The proposed measure, termed V, shares several favorable properties with an earlier V1 but also improves the handling of censoring. The statistic contrasts distance measures between individual 1/0 survival processes and fitted survival curves with and without covariate information. These distance measures, Dx and D, respectively, are themselves informative as summaries of absolute rather than relative predictive accuracy. We recommend graphical comparisons of survival curves for prognostic index groups to improve the understanding of obtained values for V, Dx, and D. Their use and interpretation is exemplified for a Yorkshire lung cancer study on survival. From this and an overview for several well-known clinical data sets, we show that the likely amount of relative or absolute predictive accuracy is often low even if there are highly significant and relatively strong prognostic factors.*

5. Stare J, Heinzl H, Harrell F: On the Use of Buckley and James Least Squares Regression for Survival Data. *New Approaches in Applied Statistics* (2000) 16:125-134;