ADEMP-PreReg
Simulation Study Plan

Project: FSGLmstate

Kaya Miah

July 5, 2024

Version: 0.1.0
Last updated: 2023-10-31
Preregistration template designed by
Björn S. Siepe, František Bartoš, Tim P. Morris, Anne-Laure Boulesteix, Daniel W. Heck, and Samuel Pawel

1 Instructions

General Information

This template can be used to plan and/or preregister Monte Carlo simulation studies according to the ADEMP framework (Morris et al., 2019). The preprint associated with this template is (Siepe et al., 2023). Alternative Google Docs and Word versions of this template are available at (https://github.com/bsiepe/ADEMP-PreReg). To time-stamp your protocol, we recommend uploading it to the Open Science Framework (https://osf.io/) or Zenodo (https://zenodo.org/). When using this template, please cite the associated preprint (Siepe et al., 2023). If you have any questions or suggestions for improving the template, please contact us via the ways described at (https://github.com/bsiepe/ADEMP-PreReg).

Using this template

Please provide detailed answers to each of the questions. If you plan to perform multiple simulation studies within the same project, you can either register them separately or number your answers to each question with an indicator for each study. As the planning and execution of simulation studies often involves considerable complexity and unknowns, it may be difficult to answer all the questions in this template or some changes may be made along the analysis pathway. This is to be expected and should not deter from preregistering a simulation study; rather, any modifications to the protocol should simply be reported transparently along with a justification, which will ultimately add credibility to your research. Finally, the template can also be used as a blueprint for the reporting of non-preregistered simulation studies.

2 General Information

2.1 What is the title of the project?

Answer

Variable selection via fused sparse-group lasso penalized multi-state models incorporating molecular data

2.2 Who are the current and future project contributors?

Answer

Kaya Miah, Jelle J. Goeman, Hein Putter, Annette Kopp-Schneider, Axel Benner

2.3 Provide a description of the project.

Explanation: This can also include empirical examples that will be analyzed within the same project, especially if the analysis depends on the results of the simulation.

Answer

We will investigate effective multi-state modeling strategies to determine an optimal, ideally parsimonious model. In particular, linking covariate effects across transitions is required to conduct joint variable selection. A useful technique to reduce model complexity is to address homogeneous covariate effects for distinct transitions based on a reparametrized model formulation. We integrate this approach to data-driven variable selection by extended regularization methods within multi-state model building. We propose the fused sparse-group lasso (FSGL) penalized Cox-type regression in the framework of multi-state models combining the penalization concepts of pairwise differences of covariate effects along with transition grouping. For optimization, we adapt the alternating direction method of multipliers (ADMM) algorithm to transition-specific hazards regression in the multi-state setting.

2.4 Did any of the contributors already conduct related simulation studies on this specific question?

Explanation: This includes preliminary simulations in the context of the current project.

Answer

No, we did not conduct previous simulation studies for investigating regularized multi-state model building.

3 Aims

3.1 What is the aim of the simulation study?

Explanation: The aim of a simulation study refers to the goal of the research and shapes subsequent choices. Aims are typically related to evaluating the properties of a method (or multiple methods) with respect to a particular statistical task. Possible tasks include ‘estimation’, ‘hypothesis testing’, ‘model selection’, ‘prediction’, or ‘design’. If possible, try to be specific and not merely state that the aim is to ‘investigate the performance of method X under different circumstances’.

Answer

The aim of the simulation study is to evaluate the model selection procedure based on FSGL penalized transition-specific hazards regression in terms of its ability to select a sparse model identifying relevant transition-specific and equal cross-transition effects.

4 Data-Generating Mechanism

4.1 How will the parameters for the data-generating mechanism (DGM) be specified?

Explanation: Answers include ‘parametric based on real data’, ‘parametric’, or ‘resampled’. Parametric based on real data usually refers to fitting a model to real data and using the parameters of that model to simulate new data. Parametric refers to generating data from a known model or distribution, which may be specified based on theoretical or statistical knowledge, intuition, or to test extreme values. Resampled refers to resampling data from a certain data set, in which case the true data-generating mechanism is unknown. The answer to this question may include an explanation of from which distributions (with which parameters) values are drawn, or code used to generate parameter values. If the DGM parameters are based on real data, please provide information on the data set they are based on and the model used to obtain the parameters. Also, indicate if any of the authors are already familiar with the data set, e.g., analyzed (a subset of) it.

Answer

In each simulation repetition, we generate multi-state data based on transition-specific hazard regression for $N = 500$ subjects as a nested series of competing risks experiments (Beyersmann et al., 2012) as depicted in Figure 1:

1. Individual in state $l \in \{1, \dots, K\}$ at time 0.
   - Waiting time t_0 in state l is generated with hazard $h_{l\cdot}(t) = \sum_{k=1, k \neq l}^{K} h_{lk}(t), t \geq 0$.
   - State X_{t_0} entered at this time is determined in a multinomial experiment with decision probability $h_{lk}(t_0)/h_{l\cdot}(t_0)$ on state $k, k \neq l$.
2. Individual has entered state k at time t_0.
   - Waiting time t_1 in state k is generated with hazard $h_{k\cdot}(t) = \sum_{\tilde{k}=1, \tilde{k} \neq k}^{K} h_{k\tilde{k}}(t), t \geq t_0$.
   - State $X_{t_0+t_1}$ entered at this time is determined in a multinomial experiment with decision probability $h_{k\tilde{k}}(t_0 + t_1)/h_{k\cdot}(t_0 + t_1)$ on state $\tilde{k}, \tilde{k} \neq k$.
3. Further competing risks experiments are carried out until reaching an absorbing state.

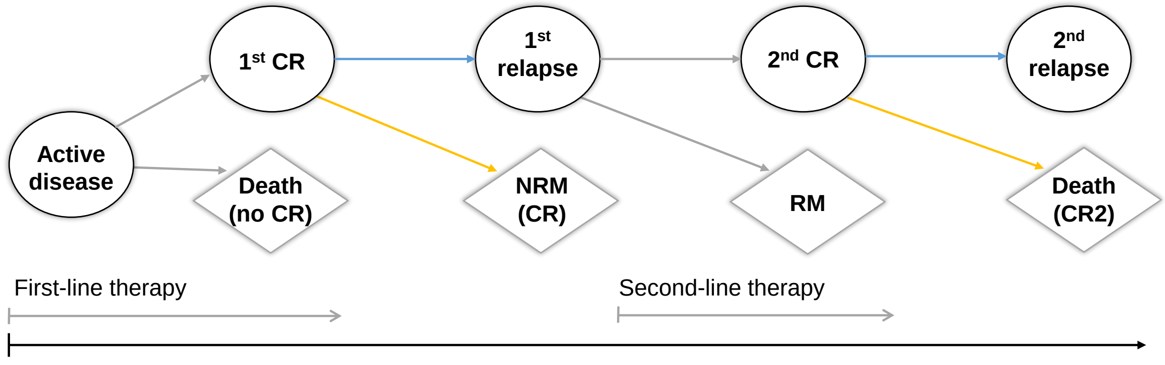

Figure 1: State chart of the multi-state model for acute myeloid leukemia (AML) with nine states and eight possible transitions represented by arrows.

4.2 What will be the different factors of the data-generating mechanism?

Explanation: A factor can be a parameter/setting/process/etc. that determines the data-generating mechanism and is varied across simulation conditions.

Answer

We will vary the following factors:

- Event times drawn from an exponential distribution as described in Section 4.1
- Design matrix X with binary covariates
- Transition-specific baseline hazards $h_{0,q}(t)$
- Regression parameter β
- Penalty parameters λ, α, γ
- Augmented Lagrangian parameter ρ
- Step size in gradient descent ϵ_{GD}
- Tolerance of stopping criterion for Cox estimation opt_{GD}
- Relative/absolute tolerance for ADMM stopping criterion $\epsilon_{rel}, \epsilon_{abs}$
- Maximum number of iterations max_{iter}

4.3 If possible, provide specific factor values for the DGM as well as additional simulation settings.

Explanation: This may include a justification of the chosen values and settings.

Answer

We will use the following values for our data-generating mechanism:

Multi-state data:

- Event times $T \sim \text{Exp}(\eta)$
- Binary covariates $X_{p,i} \sim \mathcal{B}(0.5),\ p = 1, \dots, 50, i = 1, \dots, 500$
- Transition-specific baseline hazards $h_{0,q}(t) \in \{.01, .05\}$
- Regression parameters $\beta_{p,q} \in \{-1.2, -0.8, 0, 0.8, 1.2\}$

FSGL Method:

- Penalty parameters: Optimal λ selected by generalized cross-validation (GCV); $\alpha \in \{0, 0.2, 0.5, 0.8, 1\}; \gamma \in \{0, 0.2, 0.5, 0.8, 1\}$
- Augmented Lagrangian parameter $\rho = 1$
- Step size in gradient descent $\epsilon_{GD} = 0.01$
- Tolerance of stopping criterion for Cox estimation $opt_{GD} = 10^{-6}$
- Relative/absolute tolerance for ADMM stopping criterion $\epsilon_{rel} = 10^{-2}, \epsilon_{abs} = 10^{-4}$
- Maximum number of iterations $max_{iter} = 1000$

4.4 If there is more than one factor: How will the factor levels be combined and how many simulation conditions will this create?

Explanation: Answers include ‘fully factorial’, ‘partially factorial’, ‘one-at-a-time’, or ‘scattershot’. Fully factorial designs are designs in which all possible factor combinations are considered. Partially factorial designs denote designs in which only a subset of all possible factor combinations are used. One-at-a-time designs are designs where each factor is varied while the others are kept fixed at a certain value. Scattershot designs include distinct scenarios, for example, based on parameter values from real-world data.

Answer

We will vary the conditions in a partially factorial manner, i.e. we will repeat multi-state simulations for all combinations of penalty parameters.

5 Estimands and Targets

5.1 What will be the estimands and/or targets of the simulation study?

Explanation: Please also specify if some targets are considered more important than others, i.e., if the simulation study will have primary and secondary outcomes.

Answer

Our primary target/model-based estimand focuses on the regression coefficients $\beta_{p.q}$ from the penalized Cox-type proportional hazards models

$$h_q(t|x) = h_{0,q}(t)\exp\{\beta_q^T x\},\ q = 1, \dots, Q,$$

where $h_{0,q}(t)$ denotes the baseline hazard rate of transition q at time t, $x = (x_1, \dots, x_p)^T \in \mathbb{R}^P$ the vector of covariates and $\beta_q \in \mathbb{R}^P$ the vector of transition-specific regression coefficients for P covariates.

6 Methods

6.1 How many and which methods will be included and which quantities will be extracted?

Explanation: Be as specific as possible regarding the methods that will be compared, and provide a justification for both the choice of methods and their model parameters. This can also include code which will be used to estimate the different methods or models in the simulation with all relevant model parameters. Setting different prior hyperparameters might also be regarded as using different methods. Where package defaults are used, state this. Where they are not used, state what values are used instead.

Answer

We will compare the following methods:

1. Unpenalized estimation of a multi-state model with ADMM optimization
2. Lasso penalization of a multi-state model with ADMM optimization
3. FSGL penalization of a multi-state model with ADMM optimization

7 Performance Measures

7.1 Which performance measures will be used?

Explanation: Please provide details on why they were chosen and on how these measures will be calculated. Ideally, provide formulas for the performance measures to avoid ambiguity. Some models in psychology, such as item response theory or time series models, often contain multiple parameters of interest, and their number may vary across conditions. With a large number of estimated parameters, their performance measures are often combined. If multiple estimates are aggregated, specify how this aggregation will be performed. For example, if there are multiple parameters in a particular condition, the mean of the individual biases of these parameters or the bias of each individual parameter may be reported.

Answer

1. Primary performance measure: **Sensitivity & specificity** for covariate selection

   - Mean counts of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN):

	$\#\{\beta_{p.q} \neq 0\}$	$\#\{\beta_{p.q} = 0\}$
$\#\{\hat{\beta}_{p.q} \neq 0\}$	TP	FP
$\#\{\hat{\beta}_{p.q} = 0\}$	FN	TN
	TP + FN	FP + TN

   $TPR = \frac{TP}{TP+FN}$, $TNR = \frac{TN}{FP+TN}$

2. Secondary performance measure: **Prediction accuracy**

   - Bias for non-zero predictors
   - Mean squared error (MSE) for non-zero predictors

7.2 How will Monte Carlo uncertainty of the estimated performance measures be calculated and reported?

Explanation: Ideally, Monte Carlo uncertainty can be reported in the form of Monte Carlo Standard Errors (MCSEs). Please see Siepe et al. (2023) and Morris et al. (2019) for a list of formulae to calculate the MCSE related to common performance measures, more accurate jackknife-based MCSEs are available through the `rsimsum` (Gasparini, 2018) and `simhelpers` (Joshi & Pustejovsky, 2022) R packages, the `SimDesign` (Chalmers & Adkins, 2020) R package can compute confidence intervals for performance measures via bootstrapping. Monte Carlo uncertainty can additionally be visualized using plots appropriate for illustrating variability, such as MCSE error bars, histograms, boxplots, or violin plots of performance measure estimates, if possible (e.g., bias).

Answer

We will report Monte Carlo uncertainty in tables (MCSEs next to the estimated performance measures) and in plots (error bars with ±1MCSE around estimated performance measures). We will use the formulas provided in Morris et al. (2019) to calculate MCSEs.

7.3 How many simulation repetitions will be used for each condition?

Explanation: Please also indicate whether the chosen number of simulation repetitions is based on sample size calculations, on computational constraints, rules of thumb, or any other heuristic or combination of these strategies. Formulas for sample size planning in simulation studies are provided in Siepe et al. (2023). If there is a lack of knowledge on a quantity for computing the Monte Carlo standard error (MCSE) of an estimated performance measure (e.g., the variance of the estimator is needed to compute the MCSE for the bias), pilot simulations may be needed to obtain a guess for realistic/worst-case values.

Answer

The number of simulation runs is based on the MCSE of TPR as primary performance measure of interest. Thus, we need $n_{sim} = 225$ simulation repetitions per condition as we aim for $\mathrm{MCSE}(\mathrm{TPR}) \leq 0.01$ and assume $\mathrm{MCSE}(\widehat{\mathrm{TPR}}) \leq 0.15$, resulting in $n_{sim} = \frac{0.15^2}{0.01^2} = 225$.

7.4 How will missing values due to non-convergence or other reasons be handled?

Explanation: ‘Convergence’ means that a method successfully produces the outcomes of interest (e.g., an estimate, a prediction, a *p*-value, a sample size, etc.) that are required for estimating the performance measures. Non-convergence of some iterations or whole conditions of simulation studies occurs regularly, e.g., for numerical reasons. It is possible to impute non-converged iterations, exclude all non-converged iterations or to implement mechanisms that repeat certain parts of the simulation (such as data generation or model fitting) until convergence is achieved. Further, it is important to consider at which proportion of failed iterations a whole condition will be excluded from the analysis.

Answer

We do not expect missing values or non-convergence. If we observe any non-convergence, we exclude the non-converged cases and report the number of non-converged cases per method and condition.

7.5 How do you plan on interpreting the performance measures? (optional)

Explanation: It can be specified what a ‘relevant difference’ in performance, or what ‘acceptable’ and ‘unacceptable’ levels of performance might be to avoid post-hoc interpretation of performance. Furthermore, some researchers use regression models to analyze the results of simulations and compute effect sizes for different factors, or to assess the strength of evidence for the influence of a certain factor (Chipman & Bingham, 2022; Skrondal, 2000). If such an approach will be used, please provide as many details as possible on the planned analyses.

Answer

To assess variable selection, a higher TPR and TNR of the corresponding regularization method is considered to perform better in terms of model selection. Further, we aim for little loss of predictive accuracy (i.e. smaller bias and MSE) as a secondary criterion.

8 Other

8.1 Which statistical software/packages do you plan to use?

Explanation: Likely, not all software used can be prespecified before conducting the simulation. However, the main packages used for model fitting are usually known in advance and can be listed here, ideally with version numbers.

Answer

We will use the following packages of `R` version 4.3.3 (R Core Team, 2024) in their most recent versions: The `mstate` package (Wreede et al., 2011) to generate data, `penMSM` (Sennhenn-Reulen & Kneib, 2016) to perform penalized multi-state regression, and the `ggplot2` package (Wickham, 2016) to create visualizations.

8.2 Which computational environment do you plan to use?

Explanation: Please specify the operating system and its version which you intend to use. If the study is performed on multiple machines or servers, provide information for each one of them, if possible.

Answer

We will run the simulation study on a Windows 10 machine. The complete output of `sessionInfo()` will be saved and reported in the supplementary materials.

8.3 Which other steps will you undertake to make simulation results reproducible? (optional)

Explanation: This can include sharing the code and full or intermediate results of the simulation in an open online repository. Additionally, this may include supplemental materials or interactive data visualizations, such as a shiny application.

Answer

We will upload the fully reproducible simulation script as well as all reported simulation results to GitHub (https://github.com/k-miah/FSGLmstate).

8.4 Is there anything else you want to preregister? (optional)

Explanation: For example, the answer could include the most likely obstacles in the simulation design, and the plans to overcome them.

Answer

No.

References

Chalmers, R. P., & Adkins, M. C. (2020). Writing effective and reliable Monte Carlo simulations with the SimDesign package. *The Quantitative Methods for Psychology*, *16*(4), 248–280. https://doi.org/10.20982/tqmp.16.4.p248

Chipman, H., & Bingham, D. (2022). Let's practice what we preach: Planning and interpreting simulation studies with design and analysis of experiments. *Canadian Journal of Statistics*, *50*(4), 1228–1249. https://doi.org/10.1002/cjs.11719

Gasparini, A. (2018). Rsimsum: Summarise results from Monte Carlo simulation studies. *Journal of Open Source Software*, *3*(26), 739. https://doi.org/10.21105/joss.00739

Joshi, M., & Pustejovsky, J. (2022). *Simhelpers: Helper functions for simulation studies* [R package version 0.1.2]. https://CRAN.R-project.org/package=simhelpers

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. https://doi.org/10.1002/sim.8086

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Sennhenn-Reulen, H., & Kneib, T. (2016). Structured fusion lasso penalized multi-state models. *Statistics in Medicine*, *35*(25), 4637–4659. https://doi.org/10.1002/sim.7017

Siepe, B. S., Bartoš, F., Morris, T. P., Boulesteix, A.-L., Heck, D., & Pawel, S. (2023). Simulation studies for methodological research in psychology: A standardized template for planning, preregistration, and reporting [Preprint]. https://doi.org/10.31234/osf.io/ufgy6

Skrondal, A. (2000). Design and analysis of Monte Carlo experiments: Attacking the conventional wisdom. *Multivariate Behavioral Research*, *35*(2), 137–167. https://doi.org/10.1207/s15327906mbr3502_1

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved October 17, 2023, from https://ggplot2.tidyverse.org

Wreede, L. C. d., Fiocco, M., & Putter, H. (2011). Mstate: An R Package for the Analysis of Competing Risks and Multi-State Models. *Journal of Statistical Software*, *38*(1), 1–30. https://doi.org/10.18637/jss.v038.i07