A global test for competing risks survival analysis

Edelmann, D.; Saadati, M.; Putter, H.; Goeman, J.

Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2020, Vol. 29(12) 3666–3683

DOI: 10.1177/0962280220938402
journals.sagepub.com/home/smm
SAGE

A global test for competing risks survival analysis

Dominic Edelmann[1], Maral Saadati[1], Hein Putter[2] and Jelle Goeman[2]

Abstract

Standard tests for the Cox model, such as the likelihood ratio test or the Wald test, do not perform well in situations, where the number of covariates is substantially higher than the number of observed events. This issue is perpetuated in competing risks settings, where the number of observed occurrences for each event type is usually rather small. Yet, appropriate testing methodology for competing risks survival analysis with few events per variable is missing. In this article, we show how to extend the global test for survival by Goeman et al. to competing risks and multistate models [Per journal style, abstracts should not have reference citations. Therefore, can you kindly delete this reference citation.]. Conducting detailed simulation studies, we show that both for type I error control and for power, the novel test outperforms the likelihood ratio test and the Wald test based on the cause-specific hazards model in settings where the number of events is small compared to the number of covariates. The benefit of the global tests for competing risks survival analysis and multistate models is further demonstrated in real data examples of cancer patients from the European Society for Blood and Marrow Transplantation.

Keywords

Competing risks, global test, survival, cause-specific hazards, stratified Cox model

1 Introduction

It is well known that inference in the Cox proportional hazards model based on maximization of the partial likelihood does not perform well when the number of events is small compared to the number of predictors. Under these circumstances, the parameter estimates for the log-hazard ratios can be severely biased.[1–3] Moreover, standard tests as the Wald test, the score test, and the likelihood ratio test (LRT) perform badly[1,2] and frequently do not control the specified level of significance.[3,4]

In many applications,[5–7] failures may occur due to several causes that should be treated separately in statistical modeling and testing.[8] In these competing risks settings, the association of predictor variables with different event types is typically either modeled by a cause-specific hazards[9] or a Fine and Gray model.[10] Both models involve fitting a proportional hazards model for each event type of interest. Hence, the number of events for a single cause is smaller than the total number of cases. It then often arises that there are only few events for one or more of the causes. In these settings, the Wald test, the score test, and the LRT consequently show a bad performance.

For the proportional hazards model, Goeman et al.[11] have developed a global test for testing the association of a group of p predictor variables with a (possibly right-censored) time-to-event outcome. Different from standard tests, this test is applicable in high dimensions, i.e. when the number of predictors p exceeds the sample size n.

[1]Division of Biostatistics, German Cancer Research Center, Heidelberg, Germany
[2]Biomedical Data Sciences, Leiden University, Leiden, The Netherlands

Corresponding author:
Dominic Edelmann, German Cancer Research Center, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany.
Email: dominic.edelmann@dkfz-heidelberg.de

Moreover, the global test reliably controls the specified level of significance in settings with a low number of events.

As noted above, problems arising from few events per variable (EPV) frequently show up in competing risks situations. To tackle these problems, some authors pool similar events,[12] i.e. they merge events of different types into a novel combined event type. While pooling of events may increase the number of events per event type, it leads to new problems. First, neglecting the distinctiveness of events types implies a loss of information. Second, different ways of pooling events may lead to different results in subsequent testing. When for example testing the null hypothesis of no effect of the predictors on any event type, one way of pooling might lead to acceptance, another way to rejection of the null. Such an arbitrariness of results is clearly not desirable. Yet, testing procedures for competing risks survival analysis that can deal with rare event types are missing.

The goal of this work is to develop a novel approach for testing the strong null hypothesis "no association of the predictors with any event type" in the competing risks setting that reliably controls the type I error and outperforms existing standard tests in terms of power in situations with rare event types and/or a high number of covariates.

To this end, we extend the global test for survival by Goeman et al.[11] Notably, we derive a global test for the proportional hazards model with strata based on the test established in Goeman et al.[11] Then a global test based on cause-specific hazards regression can be derived from the result for the proportional hazards model with strata. Moreover, we remark that the result for the stratified Cox model also allows us to establish a global test for general multistate models.

There are various applications of such tests in practice. First, similar to the global test for the standard Cox model,[11] the global test for the competing risks setting offers an alternative for high-dimensional data, where standard tests are not applicable (see Goeman et al.,[11] Section 5 for an example in the standard Cox model). However, we will see in Section 3.1 that there are also scenarios in the low-dimensional setting, in which the nominal type I error is much better controlled by the global test than by standard tests, and it should hence be the method of choice. Moreover, we show how to extract more detailed information from the result of global tests using closed testing procedures in Section 4.1. Finally, the global test for the multistate model (similarly for the competing risk model) offers the possibility to test if the regression coefficients for a certain subset of transitions are the same, simplifying the subsequent modeling procedure (see Section 4.2).

In a large simulation study, we investigate the properties of the global test for competing risks survival analysis. In particular, we demonstrate that the novel testing procedure reliably controls the specified level of significance in all given scenarios, including settings where the number of covariates is larger than the number of events. The power of the global test for competing risks is compared with the LRT and the Wald test in a variety of settings, and we give recommendations for statistical testing in practice. Two real data examples from the European Society for Blood and Marrow Transplantation (EBMT) illustrate the performance of the global test for competing risks survival analysis and multistate models.

The remainder of the paper is organized as follows. In Section 2.1, an extension of the global test for the proportional hazards model with strata is developed. The global test for competing risks arises as a special case, which is investigated in greater detail in Section 2.2. Notably, an alternative formula for the global test for competing risks is developed leading to a substantially faster implementation. An extension to multistate models is discussed in 2.3. Section 3 compares the performance of the global test with the LRT and the score test in a variety of simulations. After demonstrating the performance of the global test on real data examples in Section 4, Section 5 discusses the results and gives recommendations.

2 A global test for the stratified Cox model and cause-specific hazards regression

2.1 A global test for the stratified Cox model

The global test for the stratified Cox model is an extension of the global test for the ordinary Cox model derived in Goeman et al.[11] For the sake of simplicity, we will not consider the adjustment for an additional set of covariates, such as for example possible confounders. The test adjusting for additional covariates follows analogously to Goeman et al.[11] and is discussed in the supplementary material.

Let us assume that n observations of q predictors are organized in a data matrix $X \in R^{n\times q}$ with elements x_{ij}, further define $R_X = XX'$.

Moreover, we consider an additional categorical variable s with m categories and observations $s_1, \ldots, s_n$ determining the strata for each individual. The stratified Cox model, see e.g. Section 3.2 of Therneau and Grambsch,[13] models the hazard function of individual i at time t via

$$h_i(t) = h^{(s_i)}(t)\exp(r_i)$$

where $h^{(1)}(\cdot), \ldots, h^{(m)}(\cdot)$ are the unknown baseline hazards of strata $1, \ldots, m$ and $r_i = \sum_{l=1}^{q} \beta_l x_{il}$ is the linear effect of the predictors. We will assume throughout this article that both the predictors and the strata are time independent and that the censoring times are independent of the failure times given the predictors.

Observing a sample of size n consisting of the predictor matrix X, follow-up times $\mathbf{t} = (t_1, \ldots, t_n)$ and status indicators $\mathbf{d} = (d_1, \ldots, d_n)$, we are interested in testing the null hypothesis that the predictors are not associated with survival, i.e.

$$H_0 : \beta_1 = \cdots = \beta_q = 0 \tag{1}$$

without making any restriction on the number of covariates q. Notably, the test should also be valid in the high-dimensional setting, where q is larger than the sample size n. In this case, there are alternatives $\beta_1, \ldots, \beta_q$ satisfying $r_i = 0$ for all $i \in \{1, \ldots n\}$ and there is clearly no hope to detect these alternatives. Since it is not possible to establish tests that are robust against all alternatives, it appears sensible to focus our power on a set of interesting alternatives. We do so in a Bayesian fashion by putting a prior distribution on $\beta_1, \ldots, \beta_q$. In particular, we will assume that the regression coefficients $\beta_1, \ldots, \beta_q$ are random and a priori independent with mean zero and common variance τ^2. The resulting test has high power against alternatives for which large variance principal components of the data matrix X explain most of the variation in the response; some practical motivation is that small variance principal components are often related to noise, see Goeman et al.,[14] Sections 5.7 and 5.8 for details. The log-likelihood of τ^2 is then given by

$$L(\tau^2) = \log\left[\mathrm{E}_{(\beta_1,\ldots,\beta_q)}\left(\exp\left(\sum_{i=1}^{n} f_i(r_i)\right)\right)\right] \tag{2}$$

where

$$f_i(r_i) = d_i\Big(\log h^{(s_i)}(t_i) + r_i\Big) - H^{(s_i)}(t_i)\exp(r_i)$$

$H^{(k)}(t) = \int_0^t h^{(k)}(s)\mathrm{d}s$ is the cumulative baseline hazard of stratum k and the expectation is taken with respect to the distribution of $(\beta_1, \ldots, \beta_q)$.

Plugging in estimates for the baseline hazards under the null

$$\widehat{u}_i = \widehat{H}^{(s_i)}(t_i) = \sum_{t_j \le t_i} \frac{d_j 1_{\{s_j = s_i\}}}{\sum_{t_k \ge t_j} 1_{\{s_k = s_i\}}}, \quad \widehat{\mathbf{u}} = (\widehat{u}_1, \ldots, \widehat{u}_n)$$

and proceeding along the lines of Section 3 in Goeman et al.,[11] we obtain a global test statistic for the stratified Cox model given by

$$\widehat{T} = (\mathbf{d} - \widehat{\mathbf{u}})\, R_X\, (\mathbf{d} - \widehat{\mathbf{u}})' - \mathrm{trace}(R_X\, \widehat{U}) \tag{3}$$

where $\widehat{U}$ is a diagonal matrix with entries $\widehat{U}_{ii} = \widehat{u}_i$.

Testing H_0 is carried out based on a normal approximation (analogous to Goeman et al.,[11] Section 3.4) of $\widehat{T}$ using estimates of the expectation and variance. Details on the derivation of $\widehat{T}$ and corresponding estimators for the mean and variance can be found in the supplementary material.

2.2 A global test for competing risks survival analysis

In competing risks data, failure may occur due to several causes and only the first occurring failure can be observed. In these cases, the time-to-event endpoint additionally contains information about the specific cause leading to failure. As an example, consider the data from the EBMT that we will investigate in Section 4. In this scenario, death may occur due to different infections (viral, bacterial, or fungal), relapse, graft-versus-host disease (GvHD), or other causes.

Similar to testing for association between predictors and survival in classical time-to-event data, one is often interested in testing if there is an impact of the predictors on a competing risk endpoint. In the following, a global test for this problem is established.

For this purpose, we consider a competing risk setting with m different causes of interest. As in Section 2.1, the observations of the predictors are organized in a data matrix X with elements x_{ij}. For modeling the effect of the predictors on the different causes, we assume a cause-specific hazards regression,[8] i.e. the hazard function on cause k for individual i is given by

$$h_i^{(k)}(t) = h^{(k)}(t)\exp\left(r_i^{(k)}\right)$$

where $h^{(k)}(\cdot)$ is the unknown baseline hazard of cause k, $r_i^{(k)} = \sum_{l=1}^{q} \beta_l^{(k)} x_{il}$ is the linear effect of the predictors on cause k, and $\beta^{(k)}$ denotes the corresponding vector of regression coefficients. Moreover, let Z be the $nm \times nm$ block-diagonal matrix with m blocks given by

$$Z = \begin{pmatrix} X & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & X \end{pmatrix}$$

the elements of Z will be denoted by $z_{i,j}$. Defining $\mathbf{r}^{(k)} = (r_1^{(k)}, \ldots, r_n^{(k)})$, $\mathbf{r} = (\mathbf{r}^{(1)}, \ldots, \mathbf{r}^{(m)})$ and $\boldsymbol{\beta} = (\beta^{(1)}, \ldots, \beta^{(m)})$, we obtain that $r_i^{(k)}$ can be alternatively expressed as

$$r_i^{(k)} = r_{i+n(k-1)} = \sum_{l=1}^{mq} \beta z_{i,i+n(k-1)} \tag{4}$$

The purpose of the global test for competing risks is testing the null hypothesis

$$H_0 : \boldsymbol{\beta} = \left(\beta^{(1)}, \ldots, \beta^{(m)}\right) = (0, \ldots, 0) \tag{5}$$

i.e. that there is no effect of the predictors on *any* cause of interest. Analogous to Section 2.1, we assume that the components of $\boldsymbol{\beta}$, which we will denote by $\beta_1, \ldots, \beta_{mq}$ are independent and normally distributed with common variance τ^2, reducing equation (5) to

$$H_0 : \tau^2 = 0$$

Defining the status indicators for cause k by

$$d_i^{(k)} = \begin{cases} 1 & \text{if individual } i \text{ observes cause } k, \\ 0 & \text{else} \end{cases}$$

the log-likelihood of τ^2 is given by

$$L(\tau^2) = \log\left[\mathrm{E}_{\boldsymbol{\beta}}\left(\exp\left(\sum_{k=1}^{m}\sum_{i=1}^{n} f_i^{(k)}\left(r_i^{(k)}\right)\right)\right)\right] \tag{6}$$

where

$$f_i^{(k)}(r_i^{(k)}) = d_i^{(k)}\left(\log h^{(k)}(t_i) + r_i^{(k)}\right) - H^{(k)}(t_i)\exp(r_i)$$

and $H^{(k)}(t) = \int_0^t h^{(k)}(s)\mathrm{d}s$ is the cumulative baseline hazard of stratum k.

The likelihood of τ^2 in the cause-specific hazards model is just the product of the m cause-specific likelihood for all n individuals, involving a total of nm factors. More precisely, equations (2) and (6) yield that the likelihood of τ^2 in the cause-specific hazards model is the same as in a stratified proportional hazards regression including nm (pseudo-)individuals and strata $1, \ldots, m$, where the hazard function on the $i + n\,(k-1)$ th (pseudo-)individual is given by

$$h_{i+n(k-1)}(t) = h_i^{(k)}(t) = h^{(k)}(t)\exp(r_{i+n(k-1)})$$

and $r_{i+n(k-1)} = \sum_{l=1}^{mq} \beta z_{i,i+n(k-1)}$, cf. equation (4).

Hence, the global test for the cause-specific hazards regression can be traced back to the result of the global test for the stratified Cox model. To develop explicit expressions for the global test for competing risks, we define the Breslow estimate for the cumulative baseline hazard of cause k under H_0 as

$$\widehat{u}_i^{(k)} = \widehat{H}^{(k)}(t_i) = \sum_{t_j \le t_i} \frac{d_j^{(k)}}{\sum_{k=1}^{n} 1_{\{t_k \ge t_j\}}}$$

where t_i is the follow-up time of individual i. Moreover, we denote by $\mathbf{d}^{(k)} = (d_1^{(k)}, \ldots, d_n^{(k)})$ the vector containing the status indicators for cause k and $\mathbf{d} = (\mathbf{d}^{(1)}, \ldots, \mathbf{d}^{(m)})$. Analogously define $\widehat{\mathbf{u}}^{(k)} = (u_1^{(k)}, \ldots, u_n^{(k)})$ and $\widehat{\mathbf{u}} = (\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(m)})$. Applying equation (3), we obtain the test statistic

$$\widehat{T} = (\mathbf{d} - \widehat{\mathbf{u}})\, R_Z\, (\mathbf{d} - \widehat{\mathbf{u}})' - \mathrm{trace}(R_Z\, \widehat{U}) \tag{7}$$

where $R_Z = ZZ'$ and $\widehat{U}$ is a diagonal matrix whose diagonal entries are given by the elements of the vector $\widehat{\mathbf{u}}$.

Obviously, R_Z is block-diagonal with

$$R_Z = \begin{pmatrix} XX' & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & XX' \end{pmatrix}$$

Defining $R_X = XX'$ and letting $\widehat{U}^{(k)}$ denote the diagonal matrix with entries $\widehat{U}_{ii}^{(k)} = u_i^{(k)}$, equation (7) simplifies to

$$\widehat{T} = \sum_{k=1}^{m} \left[\left(\mathbf{d}^{(k)} - \widehat{\mathbf{u}}^{(k)}\right) R_X \left(\mathbf{d}^{(k)} - \widehat{\mathbf{u}}^{(k)}\right)' - \mathrm{trace}\left(R_X\, \widehat{U}^{(k)}\right) \right] = \sum_{k=1}^{m} \widehat{T}^{(k)} \tag{8}$$

where $\widehat{T}^{(k)}$ is the global test statistic in the ordinary Cox model[11] with status indicators $d_i^{(k)}$ and follow-up times t_i. We remark, that representation (8) could have also been derived directly by considering the log-likelihood of τ^2 in the cause-specific hazards model which splits into a sum of the cause-specific terms. However, we chose to derive the global test this way, because the stratified Cox model is of independent interest and can also be used to derive tests for multistate models and conditional logistic regression. Representation (8) is substantially more efficient than computing $\widehat{T}$ for the competing risks situation using equation (3), since it exploits the block structure of R_Z. Notably, the cost of computing grows linearly in m for equation (8) but quadratically for equation (3).

Estimators for the expectation and variance of $\widehat{T}$ are given by

$$\widehat{\mathrm{E}}(\widehat{T}) = \sum_{k=1}^{m} \widehat{\mathrm{E}}\left(\widehat{T}^{(k)}\right) \tag{9}$$

and

$$\widehat{\mathrm{Var}}(\widehat{T}) = \sum_{k=1}^{m} \widehat{\mathrm{Var}}\left(\widehat{T}^{(k)}\right) \tag{10}$$

where $\widehat{E}(\widehat{T}^{(k)})$ and $\widehat{\mathrm{Var}}(\widehat{T}^{(k)})$ are corresponding estimators for the expectation and variance for the standard Cox model, cf. Section 3.3 of Goeman et al.[11]

Using the same arguments as in Section 3.4 of Goeman et al.,[11] one can show that the normalized test statistic

$$\widehat{Q} = \frac{\widehat{T} - \widehat{\mathrm{E}}(\widehat{T})}{\sqrt{\widehat{\mathrm{E}}(\widehat{T})}}$$

is asymptotically standard normally distributed. This directly induces an asymptotic test for H_0. Notably, the p-value of the asymptotic test is given by $p_{\text{parametric}} = 1 - \Phi(\widehat{Q})$, where Φ is the cumulative distribution function of the standard normal distribution.

Alternatively, permutation-based testing approaches may be applied. However, it should be noted that the permutation test is only applicable under quite restrictive assumptions. First, it requires the censoring mechanism to be independent of the predictors under consideration. Second, if there are nuisance covariates such as e.g. possible confounders (see Appendix A in the supplementary material for details), the permutation-based test will not be valid. To perform a permutation version of the global test, we generate B permutations $\pi_1, \ldots, \pi_B$ of $1, \ldots, n$. For each permutation $\pi_j : \{1, \ldots, n\} \to \{1, \ldots, n\}$, we evaluate the corresponding test statistic $\widehat{T}_{\pi_j}$ based on the original set of covariates X and the permuted competing risks endpoint $\left\{(t_{\pi_j(i)}, d^1_{\pi_j(i)}, \ldots d^m_{\pi_j(i)}), i = 1 \ldots, n\right\}$. The p-value of the permutation test is then obtained by

$$p_{\text{permutation}} = \frac{\sum_{i=1}^{B} 1(\widehat{T} \leq \widehat{T}_{\pi_j})}{B+1}$$

2.3 Extension to multistate models

The ideas of Section 2.1 used to develop a global test for competing risks can also be applied to multistate models assuming that they are Markov. If the hazard rate for the transition $j \to k$ (from state j to state k) is given by

$$h_i^{(jk)}(t) = h^{(jk)}(t)\exp\left(r_i^{(jk)}\right) \tag{11}$$

where $h^{(jk)}(\cdot)$ is an unknown baseline hazard and $r_i^{(jk)} = \sum_{l=1}^{q} \beta_l^{(jk)} x_{il}$ is the linear effect of the predictors on $j \to k$, then one could consider the null hypothesis

$$H_0 : \beta_l^{(jk)} = 0 \text{ for all } (\mathrm{j}, \mathrm{k}) \in \mathcal{T}, 1 \in \{1, \ldots, q\}$$

where $\mathcal{T}$ is the set of all possible state transitions. This corresponds to testing that no predictor is associated with any of the state transitions. Assuming that the single regression coefficients $\beta_l^{(jk)}((j,k) \in \mathcal{T}, l \in \{1, \ldots, q\})$ are i.i.d. with variance τ^2 simplifies the null hypothesis to

$$H_0 : \tau^2 = 0$$

Suppose that time t in the description of the hazards refers to time since entry in state j ("clock reset" model[8]). Proceeding along the lines of Section 2.2, one can derive that a test statistic for testing H_0 in the "clock reset" multistate model is given by

$$\widehat{T} = \sum_{(j,k)\in\mathcal{T}} \widehat{T}^{(jk)}$$

where $\widehat{T}^{(jk)}$ is the global test statistic in the ordinary Cox model[11] corresponding to the transition from j to k. A global test based on the "clock-forward" model (meaning that t in equation (11) refers to the time since beginning of observation) can be derived similarly as a sum of global test statistics for delayed entry Cox models.

Another useful application of the global test is the following. A common problem in multistate models is the large number of parameters needed to describe the effect of covariates on the transitions. Often, not that many events are observed for some of the transitions, especially those which occur at the end of the chain. One strategy to deal with this problem is to assume that the effect of some covariates is the same for a certain subset of transitions $\mathcal{S}$. This makes biological sense if the subset of transitions in question are all defined by the same clinical event. By reparametrizing $\beta_l^{(jk)} = \gamma_l + \delta_l^{(jk)}$, it is then of interest to test the null hypothesis $H_0 : \delta_l^{(jk)} = 0$, for all transitions $j \to k$ in $\mathcal{S}$. A global test statistic for testing H_0 can be established along the lines of Appendix A and is given by

$$\widehat{T}_\gamma = \sum_{(j,k)\in\mathcal{S}} \widehat{T}_\gamma^{(jk)}$$

where $\widehat{T}_\gamma^{(jk)}$ is a global test statistic in a corresponding Cox model,[11] for which the estimates of the cumulative hazards are obtained by multiplying the estimated baseline hazards by $\exp\left(\sum_{l=1}^{q} \gamma_l x_{il}\right)$.

3 Simulation studies

In the following, the performance of the global test for cause-specific hazards regression is compared with that of the corresponding LRT and the Wald test. The results for the score test are very similar to the LRT and can be found in the supplementary material. Additionally to the LRT for the cause-specific hazards regression, we also consider an LRT for the ordinary Cox model, where events of all type are merged. Motivated from the setting, where the event types relate to different causes of death, we call this test the likelihood ratio test for "overall survival" (LRT-OS).

Throughout the simulations, the covariate data X follows a multivariate normal distribution with $X \sim \mathcal{N}(0, \Sigma^{(\rho)})$, where the elements of the covariance matrix $\Sigma^{(\rho)}$ are

$$\sum_{ij}^{(\rho)} = \begin{cases} 1 & \text{for } i = j, \\ \rho & \text{for } i \neq j \end{cases}$$

Except in Section 3.4, we will always assume that ρ is 0, i.e. that the covariates are pairwise independent.

The competing risks endpoint is generated using the approach of Beyersmann et al.[15] assuming a cause-specific hazards model with

$$h_i^{(k)}(t) = h^{(k)}(t)\exp(r_i^{(k)})$$

where $h^{(k)}(\cdot)$ is the unknown baseline hazard of cause k and $r_i^{(k)} = \sum_{l=1}^{q} \beta_l^{(k)} x_{il}$ is the linear effect of the predictors on cause k. The number of events k, the number of covariates q, and the cause-specific regression coefficients $\beta_1^{(k)}, \ldots, \beta_q^{(k)}$ will vary throughout the simulations. If not stated differently, the censoring times will be uniformly distributed on the interval $[0, 28]$ and independent of the predictors.

Moreover, as we will see in Section 3.1 that the type I error of the parametric versions of the tests shows stark differences, we will use permutation tests for all power comparisons. We emphasize that permutation tests will

only be valid when there are no nuisance covariates and the censoring distribution is uniform over all individuals. The purpose of using permutation tests in this simulation study is merely to obtain a fair power comparison of the different tests under consideration.

All tests will be applied with a specified significance level of $\alpha = 0.05$. Indicated empirical powers and significant levels are based on $N = 1000$ simulations. If not stated otherwise, the sample size is $n = 300$.

3.1 Comparison of type I error

All statistical tests under comparison are asymptotic tests, exploiting the fact that the distribution of the test statistics converges to a normal or chi-squared distribution. For standard tests in the Cox model, it has been demonstrated[2,3] that the asymptotic approximation for the classical tests is not satisfactory when the number of events is small compared to the number of covariates. In particular, the actual type I error of such tests if often higher than specified.

When considering competing risks models, the situation may be even more delicate since these involve fitting of one single Cox model for each event type. It is hence of crucial interest to investigate the type I error of the tests in different settings.

For this purpose, we conducted a simulation study varying the number of samples ($n = 300,\ 600$) and the number of covariates ($q = 1, 5, 10, 15, 20, 50$). For the global test, we additionally considered a high-dimensional scenario with $q = 1000$ covariates; since $n > q$, all other tests are not feasible in this case. Moreover, we considered two different censoring distributions. For the first scenario (moderate censoring), censoring is uniform on the interval $[0, 28]$ resulting in a censoring rate of approximately 33%. For the second scenario (heavy censoring), censoring is uniform on the interval $[0, 18]$ resulting in a censoring rate of approximately 54%. The baseline hazards for the two event types are equal with $h^{(k)}(\cdot) \equiv 0.05$ and the distribution of the covariates was simulated via a multivariate normal with identity covariance matrix, i.e. the covariates are pairwise independent. Since we are investigating the type I error of different tests, we assumed that the null hypothesis H_0 is true, i.e. $\beta_l^{(k)} = 0$ for $k = 1, \ldots, m$ and $l = 1, \ldots, q$. All tests were applied with a specified significance level of $\alpha = 0.05$.

Figure 1 illustrates the results of the simulations. We first note that the global test satisfyingly controls the specified level of significance in all situations, even when the number of covariates is much larger than the sample size. Like its counterparts for Cox regression[11] or the generalized linear model,[16] the global test for competing risks rather seems to be slightly conservative, in particular for situations where only few covariates are involved.

On the other hand, the rejection rate of the LRT is close to 0.05 in all scenarios with one variable but rises with an increasing number of variables. For $q = 50$ covariates, the type I error largely exceeds the nominal level of 0.05 in all scenarios under consideration. The Wald test shows a similar behavior as the LRT, however its type I error is consistently lower than that of the LRT. Somehow surprisingly, at least for uniform censoring, the Wald test seems to perform better in the scenarios with heavy censoring compared to the scenarios with moderate censoring.

To ensure reliable results for regression estimates, standard practice is to follow the one-in-ten rule proposed by Harrell et al.,[17,18] which recommends that at least 10 events per predictor covariate should be available. In the following, we will adapt the concept of EVP for investigating the question, at what point considerable problems with standard tests arise. To be precise, we consider the EPV ratios for the rarest cause, i.e. the cause with the fewest events. Adapting this idea, we remark for the four displayed scenarios that moderate problems (type I error ≥ 0.07) with the significance level of the LRT first occur at EPV values from 4.3 to 9.6 for the rarest cause. For the Wald test, the rejection rate first exceeds 0.07 at EPV values from 3.2 to 9.6 for the rarest cause.

The given results imply that the LRT and the Wald test should be used with extreme caution when the number of covariates and/or event types is high.

Due to the substantial problems of controlling the type I error shown by the LRT and the Wald test, power comparisons of the asymptotic versions of the global test with these competitors are difficult to interpret, even more so since the global test itself is rather conservative. Consequently, we used permutation versions of all competing tests for the power comparisons.

3.2 Power comparison in the situation of two different causes of failure

Most competing risks applications have only two causes of failure, usually death and some other event of interest. In cancer studies for example, one typically considers death and relapse. Similarly, for dialysis patients, competing risks may be death on dialysis and receiving a kidney transplant.[6] Due to its importance in practice, we focus on the two-event-type case first.

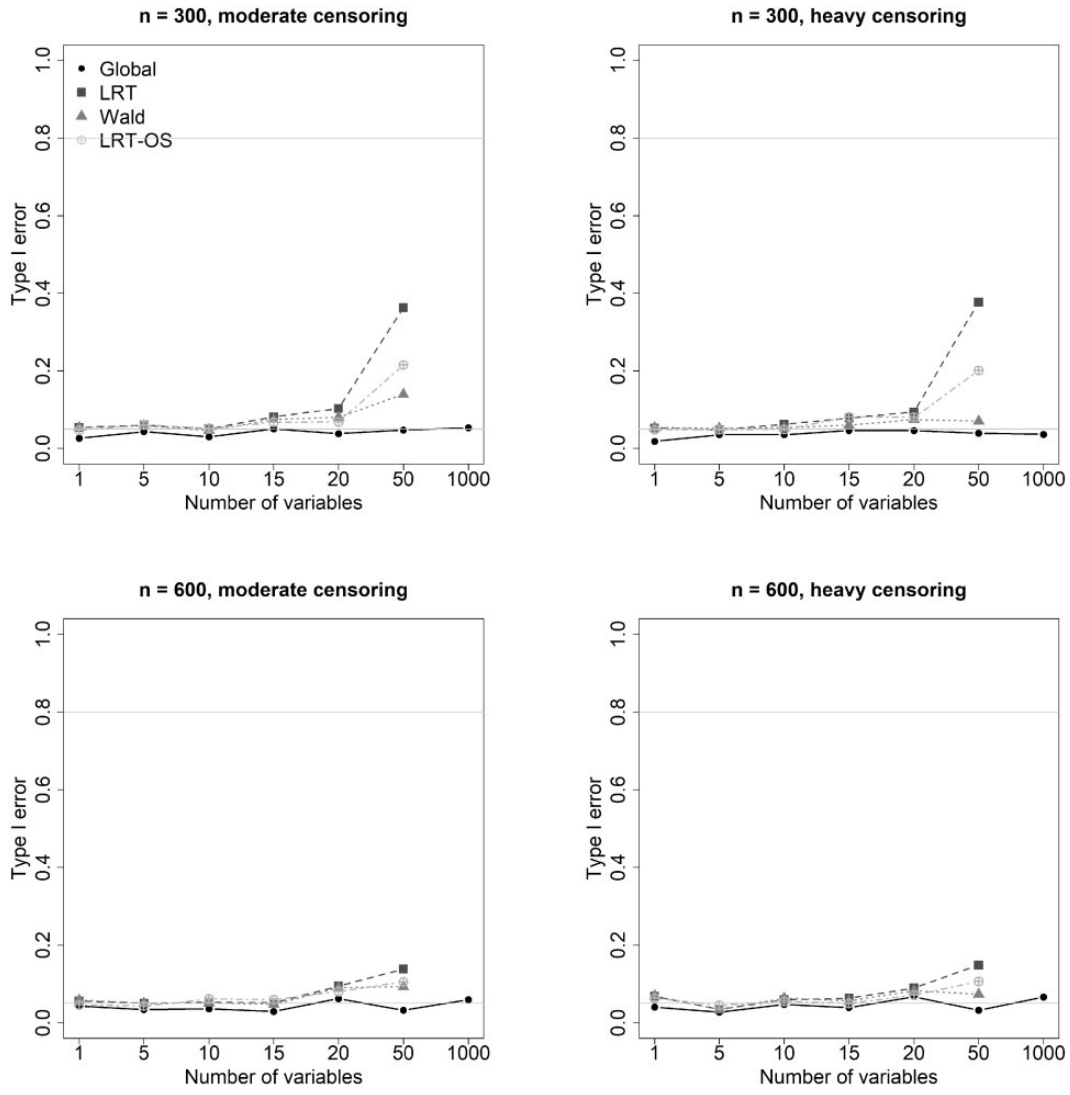

Figure 1. Empirical type I error of the global test, the LRT, the Wald test, and the LRT-OS for different numbers of variables ($q = 1, 5, 10, 15, 20, 50, 1000$) and sample sizes ($n = 300, 600$). The censoring on the left-hand side is uniform on $[0, 28]$ resulting in a censoring rate of approximately 33% over all scenarios. The censoring on the left-hand side is uniform on $[0, 18]$ resulting in a censoring rate of approximately 54% over all scenarios. The specified significance level is $\alpha = 0.05$. The results are based on $N = 1000$ simulations.

Since the actual α-level of the LRT and the Wald test can be much larger than 5%, we used permutation versions of all tests to achieve a fair comparison. Notably, we conducted a permutation version of the global test as described in Section 2.2. Permutation versions of the LRT and the Wald test are obtained analogously. Since in this simulation study there are no nuisance covariates and censoring is non-informative, permutation tests are valid. We note that in practice, we will often consider nuisance covariates or it may be reasonable to assume that the censoring mechanism depends on the covariates. Since permutation tests are not valid then, the asymptotic global test is the only test under consideration which reliably controls the specified nominal level under small EVP values. The use of permutation tests in this simulation study merely serves the purpose of providing a fair benchmark for the different tests.

In the simulations for two event types, we varied the number of covariates ($q = 2, 5, 10, 15, 20$) and the baseline hazards for the first and second cause with $(h^{(1)}(\cdot), h^{(2)}(\cdot)) \equiv ((0.05, 0.05), (0.07, 0.03), (0.09, 0.01), (0.095, 0.005))$. With decreasing baseline hazard $h^{(2)}(\cdot)$, events related to the second cause become less and less frequent, allowing

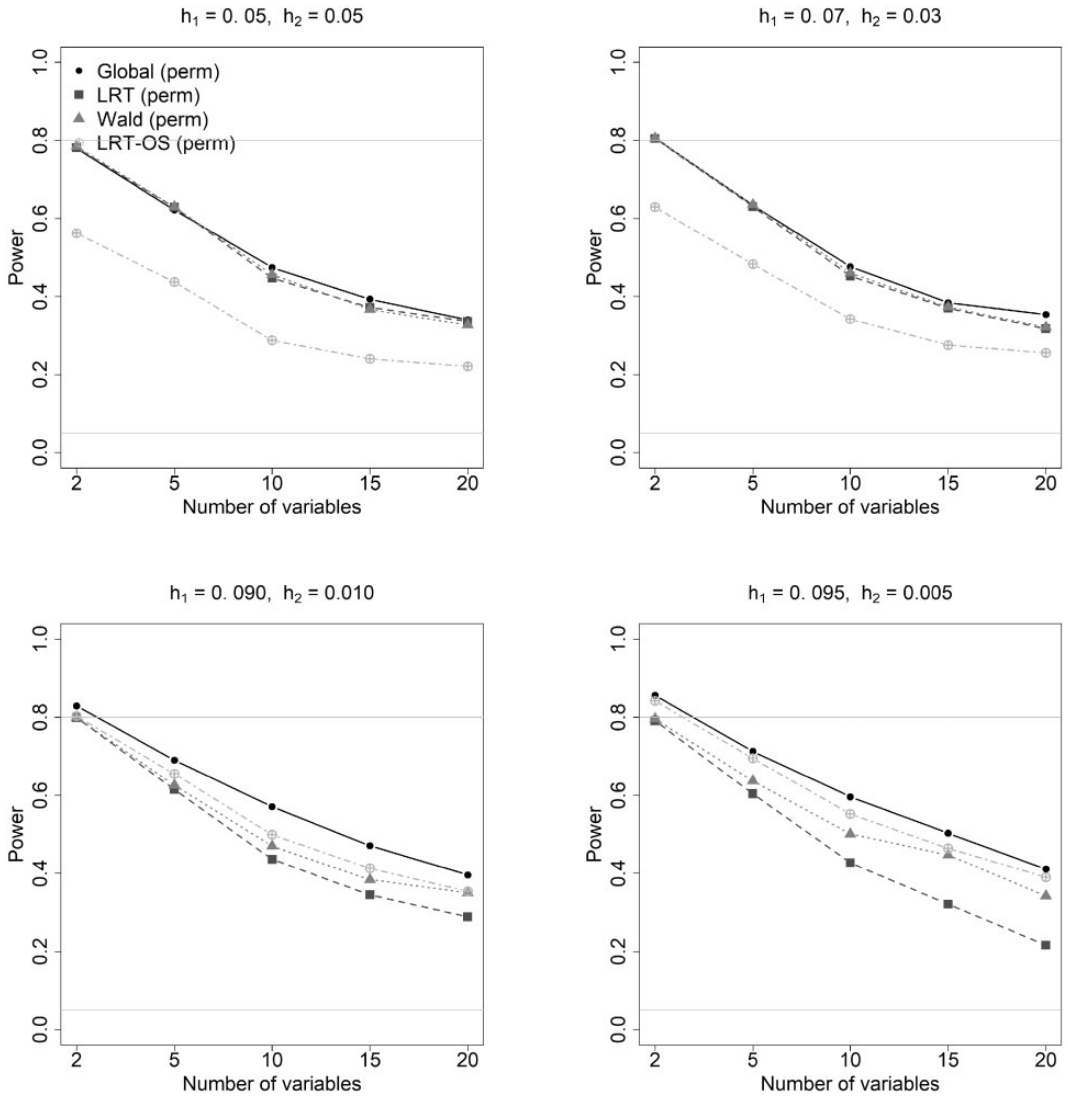

Figure 2. Power of permutation versions of the global test, the LRT, the Wald test, and the LRT-OS for different numbers of variables ($q = 2, 5, 10, 15, 20$) and two event types, where the specified significance level is $\alpha = 0.05$. The regression coefficients are given in equation (12). The results are based on $N = 1000$ simulations and a sample size of $n = 300$. The censoring times are uniformly distributed on the interval $[0, 28]$ resulting in censoring rates from 33% to 34%.

to investigate the influence of rare event types on the power of different tests. The regression coefficients on the two different causes of interest are given by

$$\beta^{(1)} = (0.25, 0, \ldots, 0)' \quad \beta^{(2)} = (0, 0.25, 0, \ldots, 0)' \tag{12}$$

Hence, the first covariate is associated with cause 1 while the second covariate is associated with cause 2. All other $q - 2$ predictors are noise variables that do not impact survival.

Figure 2 illustrates the results of the simulations. First, we note that the global test, the Wald, and the LRT show virtually the same performance for the scenario with $h^{(1)}(\cdot) = h^{(2)}(\cdot) \equiv 0.05$ The average number of events is approximately 94 resulting in EPV values for the rarest causes ranging from 1.9 ($q = 50$) to 47 ($q = 2$). The LRT for overall survival shows a substantially worse performance in this case, which is not surprising since different covariates are linked with different causes.

In situations, where one event type is rarer than the other, the global test shows slight to moderate advantages compared to the LRT and the Wald test. In particular, the power of the global test slightly increases with differences in the baseline hazards, whereas the power of the LRT decreases. The power of the Wald test decreases similarly, but the effect seems to be less severe than for the LRT. Noting that the average number of events for cause 2 is 20 for the scenario with $h^{(2)}(\cdot) \equiv 0.01$ and 10 for $h^{(2)}(\cdot) \equiv 0.005$, we see that considerable differences (>0.05) between the global test and the other two tests test first arise when the EPV is 4.07 (for the scenario with $h^{(2)}(\cdot) \equiv 0.01$ and $q = 5$ covariates) and 5.18 (for the scenario with $h^{(2)}(\cdot) \equiv 0.005$ and $q = 2$ covariates). When the number of events is close to the number of covariates, i.e. the EPV is close to 1, the differences are quite substantial, in particular between the global test and the LRT.

The performance of the LRT for overall survival gets substantially better with rarer event types. This is clearly due to the fact that overall survival is largely associated with the first covariate in these settings.

These results allow to draw interesting conclusions. First, the performance of the global test is either better than or equal to the other tests in all displayed situation, hence it uniformly dominates its competitors in these settings. Second, when one event type is much rarer than another, pooling can help to increase the power of a corresponding LRT even in situations, where different covariates are associated with different event types. Yet, we emphasize that no such pooling is necessary when applying the global test for competing risks.

3.3 Power comparison for more than two different causes of failure

In most applications, analysis is restricted to one or two different event types. This is often an oversimplification of medical reality, which is carried out for facilitating documentation and mathematical modeling. However, in the course of many diseases, a large number of interesting events can occur and modeling them separately may give deeper insights into the disease. Analyses considering a multitude of different events have been emerging in recent years[5,19] and we expect the number of corresponding studies to increase in the future.

The simulations in this subsection compare the performance of the tests under consideration for the setting of $m = 4$, 8 event types, varying the number of variables, $q = 5, 10, 15, 20$. For the baseline hazards, we considered a balanced situation where $h^{(k)}(\cdot) \equiv 0.1/m$ for all $k = 1, \ldots, m$ and an imbalanced situation, where half of the event types are rare, given by

$$h^{(k)}(\cdot) \equiv \begin{cases} 0.16/m & \text{for } k \text{ odd}, \\ 0.04/m & \text{for } k \text{ even} \end{cases} \tag{13}$$

The regression coefficients are given by

$$\beta^k = \begin{cases} (\overbrace{0, \ldots, 0}^{k-1}, 0.5, \overbrace{0, \ldots, 0}^{q-k-2})' & \text{for } k \in \left\{1, \ldots \frac{m}{2}\right\}, \\ (\overbrace{0, \ldots, 0}^{q})' & \text{for } k \in \left\{\frac{m}{2} + 1, \ldots, m\right\} \end{cases} \tag{14}$$

Hence, for $k \in \left\{1, \ldots \frac{m}{2}\right\}$, the kth cause of interest is associated with the kth covariate, whereas the causes $\left\{\frac{m}{2} + 1, \ldots, m\right\}$ are not associated with any covariate. Notably, for the case of imbalanced baseline hazards, half of the rare and half of the non-rare causes are associated with a covariate. Censoring times were again uniformly distributed on the interval $[0, 28]$ resulting in censoring rates of 33%–34%.

As in Section 3.3, we used permutation versions of the global test, LRT, and score test to achieve a fair comparison. The simulation results are illustrated in Figure 3.

The results are similar to the findings of Section 3.2. Notably, in the balanced scenarios, the global test shows only a slight or no advantage compared to the other tests; in the setting with $q = 5$ covariates and $m = 4$ balanced event types (EPV for the rarest cause 8.7), the results are virtually identical (global: 0.888, LRT: 0.888, and Wald: 0.887). In the imbalanced scenarios, however, in particular when more variables are involved, the global test clearly outperforms the LRT and the Wald test. The LRT for overall survival is not competitive in these scenarios, demonstrating that pooling gets more complicated if more event types are present.

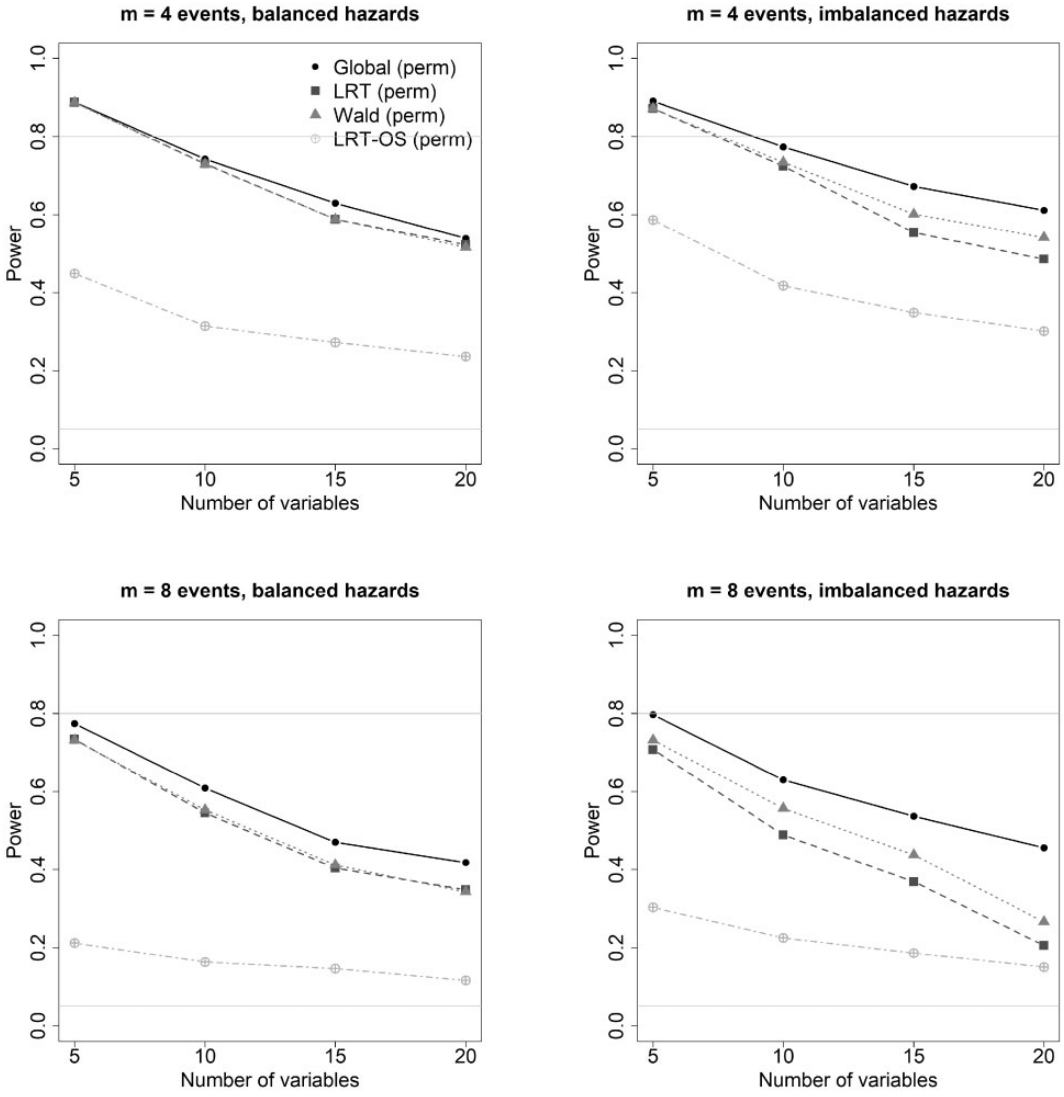

Figure 3. Power of permutation versions of the global test, the LRT, the Wald test, and the LRT-OS for different numbers of variables ($q = 5, 10, 15, 20$) and numbers of event types $m = 4, 8$, where the specified significance level is $\alpha = 0.05$. Two different scenarios are considered for the baseline hazards. In the balanced case, all baseline hazards are equal with $h^{(k)}(\cdot) \equiv 0.1/m$. For the imbalanced case, see equation (13). The regression coefficients are given in equation (14). The results are based on $N = 1000$ simulations and a sample size of $n = 300$. The censoring times are uniformly distributed on the interval $[0, 28]$ resulting in censoring rates from 32% to 33%.

Comparing the Wald test with the LRT, the impression of the previous two subsections is confirmed. While the performance of the LRT and the Wald test is virtually identical for the balanced hazards scenarios, the Wald test seems to be more robust to settings with rare events where the EVP for the rarest cause is small.

In terms of EPVs, we observe that the permutation version of the global test shows pronounced power advantages (>0.05) to the LRT and the Wald test when the EPV values for the rarest cause are between 1 and 2, e.g. in the imbalanced scenario with $m = 4$ events and $q = 15$ variables (EPV 1.2), in the imbalanced scenario with $m = 8$ events and $q = 5$ variables (EPV 1.4) or in the balanced scenario with $m = 8$ events and $q = 10$ variables (EPV 1.8)

While the differences in cases with higher EVPs appear not to be so strong, we emphasize that permutation tests are only valid when one does not correct for additional covariates in the model (e.g. possible confounders) and censoring is not uniform over all individuals. If these assumptions are violated, the null hypothesis is not exchangeable, consequently permutation versions of all tests under consideration do not control the specified level of significance.

On the other hand, we have seen in Section 3.1, that while the parametric version of the LRT and the Wald test can already show problems with type I error control for larger EPV values than 2, the parametric version of the global test reliably controls its significance level even for EPV values substantially smaller than 1.

3.4 Influence of correlation structure

In Sections 3.1–3.3, we were assuming for simplicity that the predictors are pairwise independent. In these settings, the global test outperforms its competitors if the number of EPV for the rarest cause is low. For larger EPV values, the performances of the global test, the Wald test, and the LRT are virtually identical.

When assuming correlation between the predictors[14] however, the power of the different tests under consideration can differ substantially even for high EPV values. Depending on the correlation structure of the predictors and the regression coefficients β, the global test may be more or less powerful than standard tests.

To investigates these effects, we performed simulations varying the pairwise correlation between the predictors ($\rho = 0, 0.2, 0.4, 0.6, 0.8$). We considered two different event types, where the baseline hazard for both causes is $h^{(k)}(\cdot) \equiv 0.05$. The censoring times were uniformly distributed on the interval $[0, 28]$ leading to a censoring rate of approximately 33% in all settings.

Two scenarios settings with $q = 4$ covariates were investigated. In Scenario A ("opposing effects"), the regression coefficients on causes 1 and 2 are given by

$$\beta^{(1)} = (0.25/\gamma_\rho,\ -0.25/\gamma_\rho,\ 0,\ 0)' \qquad \beta^{(2)} = (0,\ 0,\ 0.25/\gamma_\rho,\ -0.25/\gamma_\rho)' \tag{15}$$

where

$$\gamma_\rho^2 := (1,\ -1,\ 0,\ 0)\,\Sigma^{(\rho)}\,(1,\ -1,\ 0,\ 0)'$$

In Scenario B ("parallel effects"), we assumed

$$\beta^{(1)} = (0.25/\xi_\rho,\ 0.25/\xi_\rho,\ 0,\ 0)' \qquad \beta^{(2)} = (0,\ 0,\ -0.25/\xi_\rho,\ -0.25/\xi_\rho)' \tag{16}$$

with

$$\xi_\rho^2 := (1,\ 1,\ 0,\ 0)\,\Sigma^{(\rho)}\,(1,\ 1,\ 0,\ 0)'$$

The normalizing constants γ_ρ and ξ_ρ were chosen in a way such that the standard deviation of the linear predictor is always 0.25 leading to a good comparability among different scenarios and choices of ρ.

The results of the simulations are illustrated in Figure 4. While the powers of the competing tests are virtually identical for $\rho = 0$, stark differences occur when correlation is present. Notably, the Wald test and the LRT are known to be invariant under changes of correlation structure and choices of different regression coefficients. The global test on the other hand shows excellent performance in Scenario B, but completely fails for higher values of ρ in Scenario A.

The behavior of the global test for correlated covariates has been investigated in the linear model case in Goeman et al.[14] Notably, it is argued that the global test is more powerful than the F-test against alternatives for which large variance principal components of the data matrix X explain most of the variation in the response. On the other hand, if small variance components of X explain most of the variation in the response, the F-test is more powerful. Since the global test for the linear model and the global test for competing risks are similar in nature, it can be expected that the global test for competing risks shows a similar behavior.

Indeed, assuming $\rho > 0$, the principal component with the largest variance corresponds to the eigenvector

$$w = (1, 1, 1, 1)$$

of Σ. All other principal components have the same variance and correspond to the eigenspace orthogonal to w.

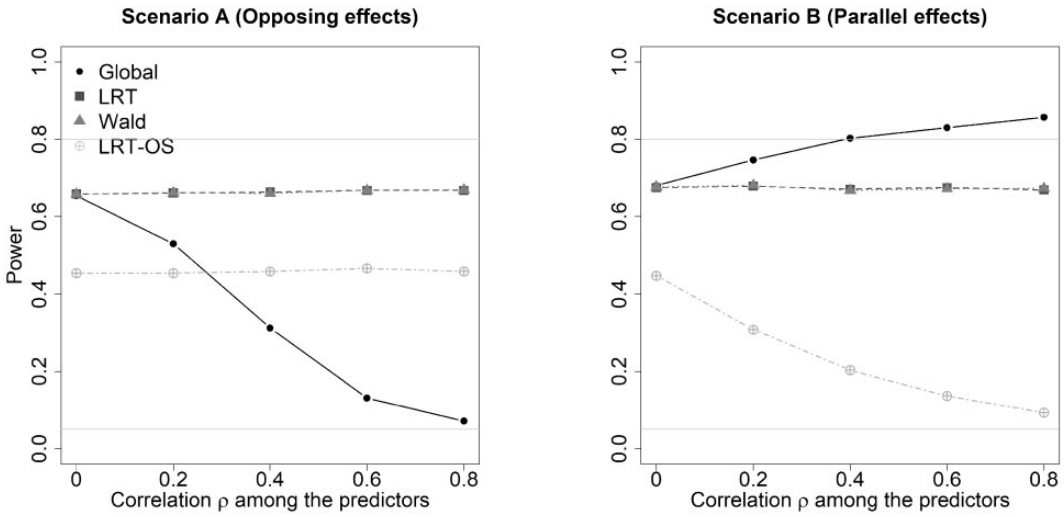

Figure 4. Power of permutation versions of the global test, the LRT, the Wald test, and the LRT-OS for two event types, where the specified significance level is $\alpha = 0.05$. We consider $q = 4$ variables and regression coefficients as given in equations (15) and (16). The results are based on $N = 1000$ simulations and a sample size of $n = 300$. The censoring times are uniformly distributed on the interval $[0, 28]$ resulting in censoring rates of approximately 33% in all settings.

While in Scenario B, the principal component corresponding to w explains a big fraction of the variation of the linear predictors for both causes, it does not explain any variation of the linear predictors in Scenario A. Consequently, in this situation, all variation is explained by small variance principal components.

The above considerations give a heuristic explanation for the differences in performance between the global test and the other tests, which are illustrated in Figure 4. In particular, one should be aware that the global test features little power against those alternatives, where a big fraction of the variation of the response can be explained by small variance principal component of the data matrix.

However, as Goeman et al.[14] note, it may be argued that small variance principal components are often dominated by uninformative noise, which can safely be assumed to be not related to the endpoint. On the other hand, the large variance principal components, i.e. the main patterns of variation in the data, are typically driven by the actual biological signal. Following this argumentation, Scenario A will occur much more rarely in practice than Scenario B. While we acknowledge that there may be some applications, where the LRT and the Wald test outperform the global test, extreme cases (such as, e.g. outlined for correlation 0.6 or 0.8 on the left-hand side of Figure 4) are rather unrealistic. We believe that this disadvantage gets clearly outweighed by the better performance of the global test for associations driven by large variance principal components. Also, we emphasize that there are situations, where the LRT and the Wald test are not feasible (if the sample size n is larger than the number of covariates q), the corresponding ML estimate does not safely converge (if the number of events is larger than q) or they do not control the nominal type I error (if the EPV ratio is small). In these situations, the global test is the method of choice.

4 Real data examples

4.1 Competing risks model

The first dataset under consideration was collected by the EBMT and comprises several thousand leukemia patients who had received bone marrow transplantation in the years from 1985 to 1998. The data are available as part of the R package mstate[20] on the Comprehensive R Archive Network.

For the purposes of this application, we restrict our analysis to all patients diagnosed with chronic myelogenous leukemia, who received transplantation between 1995 and 1998 and had non-missing T-cell depletion (TCD) status (sample size $n = 851$). The covariates in this dataset are age (categorized with levels "$\leq$20 years", "20–40 years," and ">40 years"), donor-recipient gender match ("No gender mismatch" and "Gender mismatch"), and TCD status ("No T-cell depletion" and "T-cell depletion").

Table 1. Patient characteristics for the real data example from the EBMT registry.

Age	$<=20$	65	(7.6%)
	20–40	414	(48.6%)
	>40	372	(43.7%)
Donor-recipient gender match	No gender mismatch	632	(74.3%)
	Gender mismatch	219	(25.7%)
TCD status	No T-cell depletion	754	(88.6%)
	T-cell depletion	97	(11.4%)
Cause of death	Alive	641	(75.3%)
	Relapse	36	(4.2%)
	GvHD	91	(10.7%)
	Bacterial infection	5	(0.6%)
	Viral infection	6	(0.7%)
	Fungal infection	15	(1.8%)
	Other	57	(6.7%)

EBMT, European Society for Blood and Marrow Transplantation; TCD, T-cell depletion; GvHD, graft-versus-host disease.

The following causes of death were distinguished: death due to relapse, GvHD, bacterial infection, viral infection, fungal infection, and other causes. The median follow-up time is 48.2 months. Table 1 summarizes the characteristics of the individuals in the dataset. We emphasize that the data were simplified for the purpose of illustration and no clinical conclusions should be drawn from it.

The goal of this application is to test the null hypothesis of no impact of the covariates age, donor-recipient gender match, and TCD status on any of the six different causes of death. We can recognize from Table 1 that there are several rare event types in this dataset. Notably, using dummy coding for the covariate age, the number of covariates under consideration is $q = 4$. This yields critical EPV values for the causes bacterial infection (EPV 1.25), viral infection (EPV 1.5), and fungal infection (EPV 3.75).

Since it is unlikely that they will control the nominal type I error (see Section 3.1), the LRT or Wald test should not be directly used in this situation. A typical analysis of this dataset using classical tools would now involve pooling of event types, thereby ignoring the differences between several rare causes of death. However, the simulation results in Section 3.1 yield that the global test for competing risks reliably controls the specified level of significance even in situations where there are more covariates than events.

Hence, the global test for competing risks can be directly applied on testing the null hypothesis of interest. Applying the parametric version of the global test yields a p-value of $p_{\text{parametric}} = 2.82 \times 10^{-4}$. The global test hence shows strong evidence that the covariates are associated with at least one cause of death. In particular, assuming the usual significance level of 0.05, we would reject the null hypothesis of no impact of the covariates on any cause of death.

At this point, it would be interesting for a practitioner, which causes of death are influenced by the covariates. One way to investigate this question, while attaining strict family-wise type I error control is to apply a closed testing procedure.[21] For this purpose, we consider all 2^6–1 subsets of the six different causes. The hypothesis if a subset of causes is associated with the covariates is only tested if all supersets of causes were significantly associated. This procedure also allows to calculate a multiplicity adjusted p-value[22] for each hypothesis, which is the maximum of the p-value of the test itself and of the p-values corresponding to all its supersets. Table 2 lists the multiplicity adjusted p-value for all single causes. Moreover, we also show the multiplicity adjusted p-values of the sets of causes, which were significant at a level of 0.05, but none of their subsets was.

This table provides further insight into the association of the causes and covariates. Notably, relapse is the only single cause that is significantly associated with the response. On the other hand, while GvHD itself features a p-value of slightly larger than 0.05, there seems to be some evidence that it may be associated with the covariates, since all supersets were significant. Finally, the closed testing procedure showed up a significant association of at least one of the causes bacterial infection and fungal infection with the covariates. Given the small number of events for these causes, this is a conclusion that a standard analysis could hardly provide.

4.2 Multistate model

The second data example was also collected by the EBMT and contains detailed information about the progress of 2279 leukemia patients after bone marrow transplantation. The progress of the disease of each patient can be

Table 2. Multiplicity adjusted *p*-values for single causes and significant subsets of causes.

Subsets of causes	Multiplicity adjusted *p*-value
{Relapse}	0.003
{GvHD}	0.054
{Bacterial infection}	0.079
{Viral infection}	0.064
{Fungal infection}	0.114
{Other}	0.063
{GvHD, Bacterial infection}	0.049
{GvHD, Viral infection}	0.046
{GvHD, Fungal infection}	0.033
{GvHD, Other}	0.015
{Bacterial infection, Viral infection}	0.042
{Viral infection, Other}	0.049
{Fungal infection, Other}	0.032

GvHD, graft-versus-host disease.

described by a multistate model with six different states (Transplant, Platelet Recovery (PR), Acute GvHD (AGvHD), PR and AGvHD, Relapse, and Death) and 12 possible transitions. For each patient, the covariates age ($\leq$20, 20–40, $>$40), donor-recipient gender mismatch (Yes/No), GvHD prevention (No TCD, +TCD), and year of transplant (1985–1998, 1990–1994, 1995–1998) were provided. Detailed descriptive statistics on the dataset can be found in Fiocco et al.[23] Throughout this data example, we will consider a "clock reset" model.

When modeling the impact of covariates on specific transitions, one may assume that the regression coefficients for transitions going into the same state or that are characterized by a similar clinical event are equal. On the one hand, such an assumption leads to a simplification of the multistate model, making it easier to interpret for practitioners. On the other hand, it may lead to more reliable estimates for the regression coefficients in settings where the EPV ratio for some of the transitions is small. However, it is clearly highly undesirable to make such an assumption, when the true regression coefficients differ. This urges a need for a procedure that enables testing the null hypothesis that the regression coefficients for a subset of transitions $\mathcal{S}$ are the same.

As already pointed out in Section 2.3, the global test for multistate models allows for testing this null hypothesis. For this purpose, we first fit a single stratified Cox model on the transitions in $\mathcal{S}$ assuming different baseline hazards, but the same coefficients for each transition in $\mathcal{S}$. Subsequently, we reparametrize $\beta_l^{(jk)} = \gamma_l + \delta_l^{(jk)}$, where $\beta_l^{(jk)}$ is the lth regression coefficient for the transition $j \to k$ and γ_l is the lth regression coefficient obtained from the fit of the single stratified Cox model on all transitions in $\mathcal{S}$. Considering the null hypothesis $\delta_l^{(jk)} = 0$, we can now test for equality of the regression coefficients.

In the dataset under consideration, there are four different transitions going into the state relapse (from "Transplant", "PR", "AGvHD", and "PR and AGvHD") and four different transitions going into death (from the same states that have a transition going into relapse). Testing for equality of the regression coefficients of the transitions going into relapse, we obtain a *p*-value of 0.523, an analogous test for the transitions going into death yields a *p*-value of 0.398. Both tests show no strong evidence for a violation of the assumption that the regression coefficients of the corresponding transitions are equal.

Additionally, to the sets of transitions going into the same states "Death" and "Relapse", there are two sets of transitions that are characterized by the same clinical event. Notably, these are the sets {"Transplant" $\to$ "PR", "AGvHD" $\to$ "PR and AGvHD"} (characterized by the event "PR") and {"Transplant" $\to$ "AGvHD", "PR" $\to$ "PR and AGvHD"} (characterized by the event "AGvHD"). Testing for equality of the regression coefficients of the transitions characterized by the events, PR and AGvHD yield *p*-values of 0.326 and 0.941, respectively.

In summary, no strong evidence of differing regression coefficients within the four considered subgroups of transitions could be found. It is hence an interesting option to fit four stratified Cox regressions instead of 12 different Cox regressions. Reducing the number of regression coefficients from 72 to 24, this would substantially facilitate the interpretation of the model.

5 Discussion

In situations, where the occurrence of many different event types is modeled, standard inference for competing risks models often does not perform well. For the particular problem of testing, the null hypothesis of "no effect of the predictors on any event type", we have presented evidence for the shortcomings of standard tests, such as the LRT in the cause-specific hazards model. Moreover, we have developed an alternative test for this specific null hypothesis, which does not suffer from these shortcomings. Notably, this test—the global test for competing risks analysis—reliably controls the specified level of significance in all settings considered, even when the number of covariates is larger than the number of events. Moreover, it clearly outperforms its competitors in terms of power in settings with rare event types. We emphasize again that the permutation tests used for obtaining a fair power comparison are only valid when the censoring is uninformative and there are no nuisance covariates. Since these assumptions rarely hold in practice and the parametric version of the LRT and the Wald test does not reliably control the nominal type I level, the parametric version global test is the only feasible option for small EPV values.

There are also limitations of the novel test, that we have pointed out in Section 3.4. In particular, the test has low power against alternatives for which changes in the small variance principal components explain most of the variation in the response. However, those scenarios are rather unlikely to occur in practice since small variance principal components are typically related to noise, whereas the actual signal is dominated by large variance principal components.

On the other hand, we barely touched the aspect that the global test can also be applied in the $n \gg p$ setting, where the LRT and the score test are not feasible. For high-dimensional situations, it can be expected that the global test for competing risks shows a similar performance as the related global test for survival[11] and examples of the application of this test on high-dimensional molecular data are given in Goeman et al.[11]

We would like to emphasize that in order to derive the global test for competing risks, we established a general global test for the stratified Cox model, which can be applied in many settings outside of the area of competing risks. As an important example, we have shown in Section 2.3 that a global test for general multistate models can be derived using the results from Section 2.1. The extension for multistate models enables testing the null hypothesis that the effects of some covariates are the same for a subset of transitions. This is useful in applications, where such an assumption is often made to reduce the number of parameters and we have given an example for such an application in Section 4.2. In addition, a global test for the conditional logistic regression model arises as another important special case of the global test for the stratified Cox model.

The focus of this work was on the low-dimensional competing risks setting, where we showed that there are many situations where the global test for competing risks should replace the LRT for cause-specific hazards as a standard method. However, the shortcomings of inference based on maximization of the partial likelihood do not solely affect testing of the global null hypothesis. Apart from that, important issues in practice are parameter estimation and testing for the effect of single predictors. Evaluating the potential of adapting high-dimensional methodology such as regularized regression[24,25] to tackle these issues may be a promising direction for further research.

Acknowledgements

The European Society for Blood and Marrow Transplantation (EBMT) is gratefully acknowledged for making available the data studied in Section 4.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: DE was supported by grant DFG 417754611. JG was supported by NWO VIDI grant 639.072.412.

ORCID iDs

Dominic Edelmann https://orcid.org/0000-0001-7467-6343
Hein Putter https://orcid.org/0000-0001-5395-1422

Supplemental material

Supplementary material for this article is available online.

References

1. Ogundimu EO, Altman DG and Collins GS. Adequate sample size for developing prediction models is not simply related to events per variable. *J Clin Epidemiol* 2016; **76**: 175–182.
2. Peduzzi P, Concato J, Feinstein AR, et al. Importance of events per independent variable in proportional hazards regression analysis II. Accuracy and precision of regression estimates. *J Clin Epidemiol* 1995; **48**: 1503–1510.
3. Vittinghoff E and McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. *Am J Epidemiol* 2007; **165**: 710–718.
4. Devarajan K and Ebrahimi N. Testing for covariate effect in the Cox proportional hazards regression model. *Commun Stat Theory Methods* 2009; **38**: 2333–2347.
5. D'Amico G, Morabito A, D'Amico M, et al. Clinical states of cirrhosis and competing risks. *J Hepatol* 2018; **68**: 563–576.
6. Noordzij M, Leffondré K, van Stralen KJ, et al. When do we need competing risks methods for survival analysis in nephrology? *Nephrol Dial Transplant* 2013; **28**: 2670–2677.
7. Puddu PE, Amaduzzi PL and Ricci B. Coronary heart disease incidence and competing risks: an important issue. *J Geriatr Cardiol* 2017; **14**: 425–429.
8. Putter H, Fiocco M and Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. *Stat Med* 2007; **26**: 2389–2430.
9. Holt JD. Competing risk analyses with special reference to matched pair experiments. *Biometrika* 1978; **65**: 159–165.
10. Fine JP and Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *J Am Stat Assoc* 1999; **94**: 496–509.
11. Goeman JJ, Oosting J, Cleton-Jansen AM et al. Testing association of a pathway with survival using gene expression data. *Bioinformatics* 2005; **21**: 1950–1957.
12. Franke S and Kulu H. Cause-specific mortality by partnership status: simultaneous analysis using longitudinal data from England and Wales. *J Epidemiol Community Health* 2018; **72**: 838–844.
13. Therneau TM and Grambsch PM. *Modeling survival data: extending the Cox model*. New York: Springer Science & Business Media, 2013.
14. Goeman JJ, Van De Geer SA and Van Houwelingen HC. Testing against a high dimensional alternative. *J R Stat Soc Series B Stat Methodol* 2006; **68**: 477–493.
15. Beyersmann J, Allignol A and Schumacher M. *Competing risks and multistate models with R*. New York: Springer Science & Business Media, 2011.
16. Goeman JJ, Van De Geer SA, De Kort F, et al. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* 2004; **20**: 93–99.
17. Harrell F, Lee KL, Matchar DB, et al. Regression models for prognostic prediction: advantages, problems, and suggested solutions. *Cancer Treat Rep* 1985; **69**: 1071–1077.
18. Harrell FE, Lee KL and Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996; **15**: 361–387.
19. Teixeira L, Rodrigues A, Carvalho MJ, et al. Modelling competing risks in nephrology research: an example in peritoneal dialysis. *BMC Nephrol* 2013; **14**: 110.
20. de Wreede LC, Fiocco M and Putter H. mstate: an R package for the analysis of competing risks and multi-state models. *J Stat Softw* 2011; **38**: 1–30.
21. Marcus R, Eric P and Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 1976; **63**: 655–660.
22. Dudoit S, Shaffer JP and Boldrick JC. Multiple hypothesis testing in microarray experiments. *Stat Sci* 2003; **18**: 71–103.
23. Fiocco M, Putter H and van Houwelingen HC. Reduced-rank proportional hazards regression and simulation-based prediction for multi-state models. *Stat Med* 2008; **27**: 4340–4358.
24. Ambrogi F and Scheike TH. Penalized estimation for competing risks regression with applications to high-dimensional covariates. *Biostatistics* 2016; **17**: 708–721.
25. Saadati M, Beyersmann J, Kopp-Schneider A, et al. Prediction accuracy and variable selection for penalized cause-specific hazards models. *Biom J* 2018; **60**: 288–306.