technische universität
dortmund

Seminar: Survival Analysis

Permutation Test Based on the Restricted Mean Survival Time in General Factorial Designs

Robin Grugel

July 27, 2020

Supervisor:

Dr. Marc Ditzhaus
Chair of Mathematical Statistics and Applications in Industry
Faculty of Statistics
TU Dortmund University

Contents

1. Motivation **1**

2. Methodology **1**
2.1. Factorial Designs 1
2.2. Restricted Mean Survival Time 2
2.3. Possible Hypotheses in Factorial Designs 4
2.4. Asymptotic Test for Factorial Designs 5
2.5. Permutation Test 5

3. Simulation Study **6**
3.1. Simulation Methodology 6
3.2. Simulations under $\mathcal{H}_0$: Type 1 Error Rate 7
3.2.1. Estimated Type 1 Error Rate 10
3.3. Simulations under $\mathcal{H}_1$: Power 12
3.3.1. Estimated Power 13

4. Real Data Example **14**

5. Discussion **17**

A. Additional Figures and Tables **18**

References **19**

1. Motivation

Given a setting of time to event data, especially in medical sciences, the Hazard Ratio (HR) is used to describe the effects of treatments on different groups of individuals or objects. It is the naturally incorporated effect measure already proposed by Cox (1972) as central part of the Cox Proportional Hazards Regression Model. The HR, although easily estimated in the setting of Cox models, only reveals relative effects, which lacks interpretability in clinically meaningful ways as described by Tian et al. (2018). A specific HR cannot by seen as increasing or decreasing event time rate, since the hazard function has no value in sense of actual probability. Additionally, a specification when HR's yield a practically substantial difference between clinical groups, is not given and does not lead to meaningful inference. Further missing the assumption of proportional hazard functions makes the HR hard to interpret, as mentioned by Kalbfleisch and Prentice (1981) and Lin and Wei (1989). The Restricted Mean Survival Time (RMST) offers an absolute effect measure independent from assumptions of proportionality regarding the hazards of studied clinical groups. It gives the expected time till the event occurs. In a practical example of a study with a control and a treatment group an effect can be analysed by the difference of the RMSTs as proposed in Royston and Parmar (2013) using testing procedures. As the simple comparison of two groups is often not satisfactory in complex study designs, the desired approach involves adjustments to fit general factorial designs. Those designs are commonly used in medicine or agricultural science and often bring along the characteristic of small sample sizes in all groups of the design. Approaching this setting, a asymptotic hypothesis test using a Wald type statistic for the RMST is constructed and adapted for frequently occurring small sample sizes using the resampling method of random permutations. In detailed simulations, the type 1 error rate and the power is evaluated for typical survival time distributions including the most interesting case of crossing hazard functions. Finally an example for real survival data illustrates the easy use of the implemented procedure using a clinical trial on adjuvant chemotherapy, treating colon cancer. It can clearly be seen, that the type 1 error rate of the asymptotic test is substantially inflated for small sample sizes, high censoring rates and unbalanced designs, while random permutation based testing seems nearly unaffected by disturbing influences. In the following report the theoretical basis of survival properties, factorial designs, as well as the asymptotic test and its permutation counterpart will be presented. An extensive presentation of simulation results and a critical discussion is followed by the before mentioned data example.

2. Methodology

2.1. Factorial Designs

Using General factorial designs in various fields of research aims at different statistical and practical problems. Gathering data for a specific scientific research goal can be highly time demanding and particularly expensive, therefore it is desired to collect data just in the amount and structure necessary to be able to come to conclusions for hypotheses. An easy example is the origin of factorial designs in agricultural research

already investigated by Fisher (1992), where the cultivation of fields is both costly and time consuming. To derive valid inference which crop and fertilizer leads to the highest yield a carefully controlled design is necessary. In a survival context the need for thoroughly controlled designs can be illustrated by researching expensive medication for a rare disease in pharmacology. To address this in a statistical manner a set of methods (e.g. latin hyper cubes, completely randomized designs, Plackett-Burman design etc.) is known, to ensure valid statistical inference and can be further investigated by Montgomery (2013). The amount of individuals, here for different factors and their levels are chosen carefully to be balanced. Advantages of factorial designs are the possibility to infer on main and interaction effects simultaneously. The reproducibility of experiments allows more powerful comparison of differing studies. A simple example for a factorial design is a (2×3)-design, although easily extended to arbitrary many factors and levels. For later simulations and examples, table 1 illustrates a design for factors A with levels $i_A = 1, 2$ and B with levels $i_B = 1, 2, 3$.

i_A / i_B	1	2	3
1	$(1,1)$	$(1,2)$	$(1,3)$
2	$(2,1)$	$(2,2)$	$(2,3)$

Table 1: Simple (2×3)-design with factor combinations (i_A, i_B)

Similar to Ditzhaus, Fried, et al. (2019), the goal is to utilize the factorial design for significance tests, but now in a time-to-event data setting.

2.2. Restricted Mean Survival Time

To simplify notation, consider the independent survival time random variables $T_{ij} \sim S_i$ and independent from them, censoring time random variable $C_{ij} \sim G_i$ for group index $i = 1, ..., k$ and index $j = 1, ..., n_i$ for the group specific sample sizes. A random censoring mechanism is assumed to be present. This leads to the observable random variable $X_{ij} = \min(T_{ij}, C_{ij})$ with event indicator $\delta_{ij} = \mathbb{1}_{\{X_{ij}=T_{ij}\}}$. The before mentioned factors A with $a = 2$ levels and B with $b = 3$ levels, lead to $k = a \cdot b$ subgroups with index $i = (i_A, i_B)$, which are presented in table 2.

(i_A, i_B)	$(1,1)$	$(1,2)$	$(1,3)$	$(2,1)$	$(2,2)$	$(2,3)$
i	1	2	3	4	5	6

Table 2: Notation for simple (2×3)-design with factor combinations (i_A, i_B)

The presented random variables are not necessarily continuous, but simulations will be conducted for the non discrete case only.
In contrast to using the hazard rate as an effect measure, the restricted mean survival time μ_i of a random variable $T_{i1} = T_i$ (, as i.i.d. event times within the i-th distribution) limited by a predefined time point τ by the survival time $X_i = \min(T_i, C_i)$ is defined as the area under the survival function $S(t)$, supported on the interval $(0, \tau)$ defined

by Royston and Parmar (2013) as

$$\mu_i = \mathrm{E}(\min(T_i, \tau)) = \int_0^\tau S_i(t)dt,$$

with $S_i(\tau) < 1$ and $G_i(\tau) > 0$ for $\tau > 0$. Choosing $\tau = \infty$ would lead to the ordinary expected value of the event time. The interpretation is clearly meaningful as the expected survival time within a time interval $(0, \tau)$, in sense of the time a patient can expect to pass until the event occurs. Compared to a simple value of the survival function $S_i(t)$, the RMST incorporates more information of the underlying distributional properties of the survival time and therefore also gathers the influence of crossing hazard rates in cases of non-proportionality, as pointed out by Zhao, Claggett, et al. (2016). A clear advantage of the RMST towards the median life time is, its meaningfulness for settings with a low hazard rate or a short study period, where less than 50% of the observed population experience an event, as pointed out by Kim et al. (2017). The interpration is especially useful in comparing different groups by using the difference

$$\Delta = \mu_v - \mu_u = \int_0^\tau S_v(t) - S_u(t)dt,$$

which was suggested by Zhao, Tian, et al. (2012). Plotting Δ for different values of τ can also reveal the relation between groups, e.g. the effect of a treatment in different phases of a disease. This can be helpful in illustrating the effect in non-proportional hazard cases A'Hern (2016) reports, e.g. where the difference is high in the beginning and decreases over time.
A possible approximation based on an ordered list of n_i event time points $X_{i(1)} < X_{i(2)} < ... < X_{i(n_i)}$ with $\tau := n_{i+1}$ and with a time discrete $S_i(t)$ in sense of a step function is intuitively given by

$$\mu_i \approx \sum_{j=1}^{n_i} (X_{i(j+1)} - X_{i(j)}) S_i(X_{i(j)}),$$

as summing up the areas with width $(X_{i(j+1)} - X_{i(j)})$ and height $S_i(X_{i(j)})$. The RMST can be estimated by replacing the survival function with an appropriate estimator, for example the Kaplan-Meier estimator defined as

$$\widehat{S}_k(t) = \prod_{j:X_{ij} \leq t} \left(1 - \frac{\delta_{ij}}{Y_i(X_{ij})}\right), \text{ for } t \geq 0$$

with counting process $Y_i(t) = \sum_{j=1}^{n_i} \mathbb{1}_{\{X_{ij} \geq t\}}$ for individuals under risk and resulting in $\widehat{\mu}_i = \int_0^\tau \widehat{S}_i(t)dt$. This can again be approximated for a finite set of time points by $\widehat{\mu}_i \approx \sum_{j=1}^{n_i} (X_{i(j+1)} - X_{i(j)}) S_i(X_{i(j)})$. The choice of more complex estimators, as the kernel based semi parametric Ramlau-Hansen estimator leads to a computationally more demanding estimation including elaborated bandwidth selection procedure and especially costlier numerical integration techniques. Therefore it is reasonable to choose the Kaplan-Meier estimator, as it already is the widely used standard method.

The variance of μ_i can be derived by using the asymptotic normality for $\sqrt{n_i}(\widehat{\mu}_i - \mu_i) \underset{n_i \to \infty}{\to} V \sim \mathrm{N}(0, \sigma_i^2)$. This lead to the integral

$$\sigma_i^2 = \int_0^\tau \left(\int_x^\tau S_i(t)dt \right)^2 \frac{\lambda_i(x)}{G_i(x)S_i(x)} dx,$$

with event time hazard rate $\lambda_i(t)$. Replacing the quantities of event time random variables with typical estimators lead to a estimator for the variance. $S_i(t)$ is estimated by the Kaplan-Meier estimator as before, $G_i(t)S_i(t)$ by $\frac{1}{n_i}Y_i(t) = \frac{1}{n_i}\sum_{j=1}^{n_i} \mathbb{1}_{\{X_{ij} \geq t\}}$ and the complete expression $\int_0^\tau h(t)\lambda_i(t)dt$ is approximated by $\sum_{j=1}^{n_i} h(X_{ij})\mathbb{1}_{\{X_{ij} \geq \tau\}} \frac{\delta_{ij}}{Y_i(X_{ij})}$. The resulting estimator

$$\widehat{\sigma}_i^2 = \sum_{j=1}^{l_i} \mathbb{1}_{\{t_{ij} \leq \tau\}} \frac{\delta_{ij}}{Y_i(t_{ij})} \left(\int_{t_{ij}}^\tau \widehat{S}_i(t)dt \right)^2 \frac{1}{n_i^{-1}Y_i(t_{ij})}$$

is valid for discrete and continuous event and censoring time distributions. $\widehat{\sigma}_i^2$ is a consistent estimator for σ_i^2 and will be used throughout the significance test procedure.

2.3. Possible Hypotheses in Factorial Designs

Similar to the description of Ditzhaus, Janssen, et al. (2020) the RMST $\mu_i := \mu_{(i_A, i_B)}(\tau)$, depending on a fixed τ can be decomposed additively in $\mu_{i_A, i_B}(\tau) = \mu_0(\tau) + \mu_{i_A}(\tau) + \mu_{i_B}(\tau) + \mu_{i_A i_B}(\tau)$. This decomposition leads to possible hypotheses, whose answers motivate the complete approach of significance testing in factorial designs. Basically a k-sample test for equality could be conducted by defining the contrast matrix $\boldsymbol{P}_k = \boldsymbol{I}_k - \frac{1}{k}\boldsymbol{J}_k$ with $\boldsymbol{I}_k$ as $k \times k$ identity matrix and $\boldsymbol{J}_k$ as $k \times k$ unity matrix. Noting the vector of RMSTs as $\boldsymbol{\mu} = (\mu_1 = \mu_2 = ... = \mu_k)'$ enables to formulate the other desired null hypotheses in the following fashion:

(a) $\mathcal{H}_0(\boldsymbol{H}) = \{\boldsymbol{H}\boldsymbol{\mu} = \boldsymbol{0}\} = \{\mu_1 = \mu_2 = ... = \mu_k\}$,
(b) $\mathcal{H}_0(\boldsymbol{H}_A) = \{\boldsymbol{H}_A\boldsymbol{\mu} = \boldsymbol{0}\} = \{\overline{\mu}_{1\bullet} = \overline{\mu}_{2\bullet} = ... = \overline{\mu}_{a\bullet}\}$,
(c) $\mathcal{H}_0(\boldsymbol{H}_B) = \{\boldsymbol{H}_B\boldsymbol{\mu} = \boldsymbol{0}\} = \{\overline{\mu}_{1\bullet} = \overline{\mu}_{2\bullet} = ... = \overline{\mu}_{b\bullet}\}$
(d) $\mathcal{H}_0(\boldsymbol{H}_{AB}) = \{\boldsymbol{H}_{AB}\boldsymbol{\mu} = \boldsymbol{0}\} = \{\mu_{i_A i_B} - \overline{\mu}_{i_A\bullet} - \overline{\mu}_{\bullet i_B} + \overline{\mu}_{\bullet\bullet} = 0\}$ for i_A and i_B.

$\boldsymbol{H} = \boldsymbol{P}_k$ $((a))$ delivers the k-sample hypotheses for equality, while contrast matrices $\boldsymbol{H}_A = \boldsymbol{P}_a \otimes \frac{1}{b}\boldsymbol{J}_b$ $((b))$, $\boldsymbol{H}_B = \frac{1}{a}\boldsymbol{J}_a \otimes \boldsymbol{P}_b$ $((c))$ and $\boldsymbol{H}_{AB} = \boldsymbol{P}_a \otimes \boldsymbol{P}_b$ $((d))$ lead to hypotheses for main and interaction effects. The dotted indices indicate the arithmetic means for the regarding factor levels. To address this hypotheses in a standard model, a Cox proportional hazards regression with additional dummy variables for the factorial levels, including interaction effects, could be fitted. For a high number of factor levels a lot of dummy variables lead to an unstable maximum partial likelihood estimation, especially for small sample sizes. Therefore a different approach is chosen.

2.4. Asymptotic Test for Factorial Designs

Defining the projection matrix $\boldsymbol{T} = \boldsymbol{H}'(\boldsymbol{H}\boldsymbol{H}')^{+}\boldsymbol{H}$ leads to a symmetric and idempotent contrast matrix, which allows numerically stable calculations while testing for the same null hypothesis as $\boldsymbol{H}$ itself. The Wald Type test statistic defined as

$$W_n(\boldsymbol{T}) = n(\boldsymbol{T}\widehat{\boldsymbol{\mu}})'(\boldsymbol{T}\widehat{\boldsymbol{\Sigma}}\boldsymbol{T}')^{+}\boldsymbol{T}\widehat{\boldsymbol{\mu}}$$

with estimated covariance matrix $\widehat{\boldsymbol{\Sigma}} = \operatorname{diag}\left(\frac{n}{n_1}\widehat{\sigma}_1^2, \frac{n}{n_2}\widehat{\sigma}_2^2, ..., \frac{n}{n_k}\widehat{\sigma}_k^2\right)$ and vector of estimated RMSTs $\widehat{\boldsymbol{\mu}} = (\widehat{\mu}_1, \widehat{\mu}_2, ..., \widehat{\mu}_k)'$. The statistic could be modified for testing more general hypotheses of $\boldsymbol{H}\boldsymbol{\mu} = \boldsymbol{d}$ by using the original term $(\boldsymbol{T}(\widehat{\boldsymbol{\mu}} - \boldsymbol{d}))$ in the calculation of statistics. Using consistent asymptotic estimators for the variance, it is necessary to assume that no group i vanishes in relation to the others:

$$0 < \liminf_{n\to\infty} \frac{n_i}{n} \leq \limsup_{n\to\infty} \frac{n_i}{n} < 1.$$

This accounts for different paces of growth in the sample sizes and leads to the asymptotic property of converging $W_n(\boldsymbol{T})$ in distribution:

$$W_n(\boldsymbol{T}) \underset{n\to\infty}{\overset{d}{\to}} \chi^2_{\operatorname{rank}(T)},$$

if $\mathcal{H}_0(\boldsymbol{T})$ holds and diverging $W_n(\boldsymbol{T}) \underset{n\to\infty}{\overset{p}{\to}} \infty$ if $\mathcal{H}_1(\boldsymbol{T})$ is present. Therefore the null hypothesis can be rejected by comparing $W_n(\boldsymbol{T})$ with the α-quantile of the $\chi^2_{\operatorname{rank}(T)}$-distribution yielding the formal test $\phi_n = \mathbb{1}_{\{W_n(\boldsymbol{T}) > \chi^2_{1-\alpha,\operatorname{rank}(T)}\}}$. The constructed test for the hypotheses, designed by using a specific $\boldsymbol{T}$, is consistent with the significance level α, formally defined as

$$\mathrm{E}(\phi_n) = P(W_n(\boldsymbol{T}) \geq \chi^2_{1-\alpha,\operatorname{rank}(T)}) \underset{n\to\infty}{\to} \mathbb{1}_{\mathcal{H}_1(\boldsymbol{T})} + \alpha\mathbb{1}_{\mathcal{H}_0(\boldsymbol{T})}.$$

2.5. Permutation Test

For small sample sizes, the asymptotic property of the test statistic converging in probability to a certain distribution, does not hold true. Addressing finite samples, this situation leads directly the widely used procedure of permutation methods. Under the restricted null hypotheses $\tilde{\mathcal{H}}_0$ gilt $S_1 = S_2 = ... = S_k$, the membership the k-th group is exchangeable and leads to asymptotic exactness of the test. The basic idea is, that if $\mathcal{H}_0$ is not true, the actually observed test statistics is more likely to be quite high and therefore lies in the upper tail of the distribution. There are only few permuted test statistics that exhibit a higher value. Based on this idea, Hemerik and Goeman (2017) recommend a simple estimate for the p-value by calculating the proportion of permuted statistics greater than the observed one. Because the number of all possible permutations exceeds computational feasibility, the method of random permutations is used. The result are n_{perm} random permuted data sets $\{(X^{\pi}_{ij}, \delta^{\pi}_{ij})\ \forall i = 1, ..., k \text{ and } j = 1, ..., n_i\}$. For every data set the studentized statistic as above

$$W^{\pi}_n(\boldsymbol{T}) = n(\boldsymbol{T}\widehat{\boldsymbol{\mu}}^{\pi})'(\boldsymbol{T}\widehat{\boldsymbol{\Sigma}}^{\pi}\boldsymbol{T}')^{+}\boldsymbol{T}\widehat{\boldsymbol{\mu}}^{\pi}$$

with $\widehat{\mathbf{\Sigma}}^{\pi} = \text{diag}\left(\frac{n}{n_1}\widehat{\sigma}_1^{\pi 2}, \frac{n}{n_2}\widehat{\sigma}_2^{\pi 2}, ..., \frac{n}{n_k}\widehat{\sigma}_k^{\pi 2}\right)$ and $\widehat{\boldsymbol{\mu}}^{\pi} = (\widehat{\mu}_1^{\pi}, \widehat{\mu}_2^{\pi}, ..., \widehat{\mu}_k^{\pi})'$ is calculated. The actual test decision is made by replacing the $\chi^2_{\alpha,\text{rank}(T)}$-quantile by the empirical $(1-\alpha)$-quantile of the permuted test statistic $c^{\pi}_{1-\alpha}$. This yields the formal statistical test $\phi_n^{\pi} = \mathbb{1}_{\{W_n^{\pi}(\boldsymbol{T}) > c^{\pi}_{1-\alpha}\}}$. For $n \to \infty$, $W_n^{\pi}(\boldsymbol{T})$ converges in probability under $\mathcal{H}_0$ and $\mathcal{H}_1$ as follows:

$$W_n^{\pi}(\boldsymbol{T}) \underset{n\to\infty}{\overset{p}{\to}} \chi^2_{\text{rank}(T)}.$$

This property is easily illustrated in figure 1 for two different sample sizes.

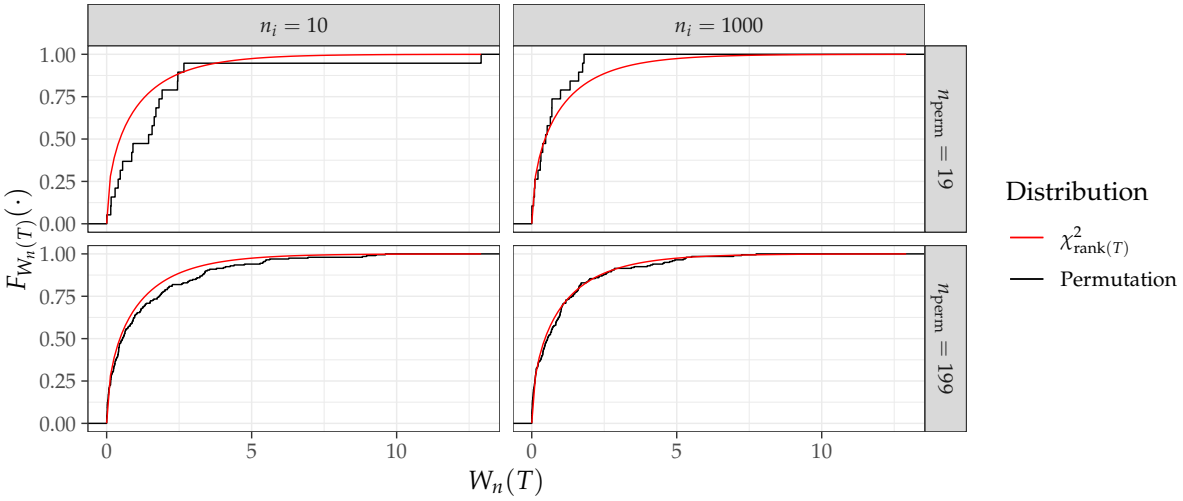

Figure 1: Asymptotic and Permutation distributions of the Wald Type test statistic

3. Simulation Study

All procedures and simulations were implemented in `R` by R Core Team (2019). The `R`-Code for all implemented tests an simulations can be found on Github[1], with extensive documentation. The widely known `survival` package by Terry M. Therneau and Patricia M. Grambsch (2000) was used to conduct simple estimation, while `dplyr` by Wickham, François, et al. (2020), `purrr` by Henry and Wickham (2020) and `tidyr` by Wickham and Henry (2020) were used to maintain a consistent programming style. `ggplot2` by Wickham (2016), `survminer` by Kassambara et al. (2020), `tables` by Murdoch (2020) and `xtable` by Dahl et al. (2019) ensure a simple and comprehensible presentation of results.

3.1. Simulation Methodology

To investigate the properties of both significance tests further and compare them especially in finite sample cases, a simulation study for the asymptotic and the permutation test is carried out with a variety of settings with focus on practical scenarios. In evaluating the properties of a statistical significance test ϕ_n all possible decisions are illustrated in table 3, which also highlights the two important properties, to wrongfully

[1] `https://github.com/robingrugel/RMST-Factorial-Design-Test`

reject $\mathcal{H}_0$ (type 1 error) and correctly reject the null hypothesis, if $\mathcal{H}_1$ is actually true (power).

	$\mathcal{H}_0$ is true	$\mathcal{H}_1$ is true
$\phi_n = 1$	type 1 error	right decision
$\phi_n = 0$	right decision	type 2 error

Table 3: possible decision in statistical significance testing

To obtain those propertiesm, datahas to be simulated under $\mathcal{H}_0(\boldsymbol{T})$ and $\mathcal{H}_1(\boldsymbol{T})$ to estimate $\mathrm{E}(\phi_n)$. In the following $n_{sim} = 5000$ simulation for each scenario and the significance level $\alpha = 0.05$ were selected to estimate the values by calculating $\widehat{\mathrm{E}}(\phi_n) = \frac{1}{n_{\text{sim}}} \sum_{l=1}^{n_{\text{sim}}} \mathbb{1}_{\{p_l < \alpha\}}$. The number of rejected null hypotheses $\sum_{l=1}^{n_{\text{sim}}} \mathbb{1}_{\{p_l < \alpha\}} =: Z$ can be seen as a sum of Bernoulli experiments, so $Z \sim \text{Bin}(n_{\text{sim}}, p_0 = \alpha)$. To decide, if ϕ_n is consistent and the level α is asymptotically met, the use of the binomial confidence interval is indicated. Using the normal approximation leads to the limits

$$p_{\text{lower/upper}} = p_0 \pm u_{1-\frac{\alpha}{2}} \sqrt{\frac{p_0(1-p_0)}{n_{\text{sim}}}},$$

with the standard normal $\left(1 - \frac{\alpha}{2}\right)$-quantile $u_{1-\frac{\alpha}{2}}$. Using the defined specification leads to $CI = (p_{\text{lower}}, p_{\text{upper}}) = (0.044, 0.056)$, which will be used to evaluate the simulated values for the type 1 error rate. Values above the upper bound indicate a too liberal decision, while values under the lower bound suggest a too conservative decision. The goal is a test which exhibits an type 1 error rate, which is covered by the confidence interval.

3.2. Simulations under $\mathcal{H}_0$: Type 1 Error Rate

Considering a (2×3)-design for the two factors A and B representative for a common use case for factorial designs gives 6 groups defined by unique combinations of factor levels. We include different sampling designs as completely balanced samples $n_{\text{bal}} = \{10, 10, 10, 10, 10, 10\}$ could be potentially rare in practical applications, especially in a survival context. Therefore an unbalanced design $n_{\text{unbal}} = \{8, 12, 14, 10, 8, 8\}$ is added. To evaluate the asymptotic property of both tests, the factor $q = 1, 2, 4$ is defined, which increases the size of the before mentioned sampling design sizes $q \cdot n_*$ multiplicative for $* = \text{bal}, \text{unbal}$ and $q = 1, 2, 4$. The influence of the censoring rate is especially relevant for small sample sizes. Therefore different but fixed rates are defined from low censoring $c_{\text{low}} = \{8\%, 7\%, 10\%, 9\%, 6\%, 6\%\}$ to intermediate $c_{\text{mid}} = \{16\%, 25\%, 21\%, 20\%, 15\%, 23\%\}$ up to high censoring $c_{\text{high}} = \{41\%, 45\%, 29\%, 35\%, 40\%, 34\%\}$. To secure, that the censoring rate is approximately equal for all n_{sim} simulated data sets, a special censoring mechanism is specified later. The tested hypothesis are for one main effect $\boldsymbol{H}_A$ representing tests for all main effects and the interaction effect $\boldsymbol{H}_{AB}$. For the simple case of proportional (and in our case even equal) curves, common continuous event time distributions are used. Figure 2 illustrates three event time distributions, $\text{Exp}(1)$, $\text{LogN}(0, 0.25)$ and $\text{Weibull}(0.4, 0.8)$, which exhibit clearly different shapes of survival probabilities.

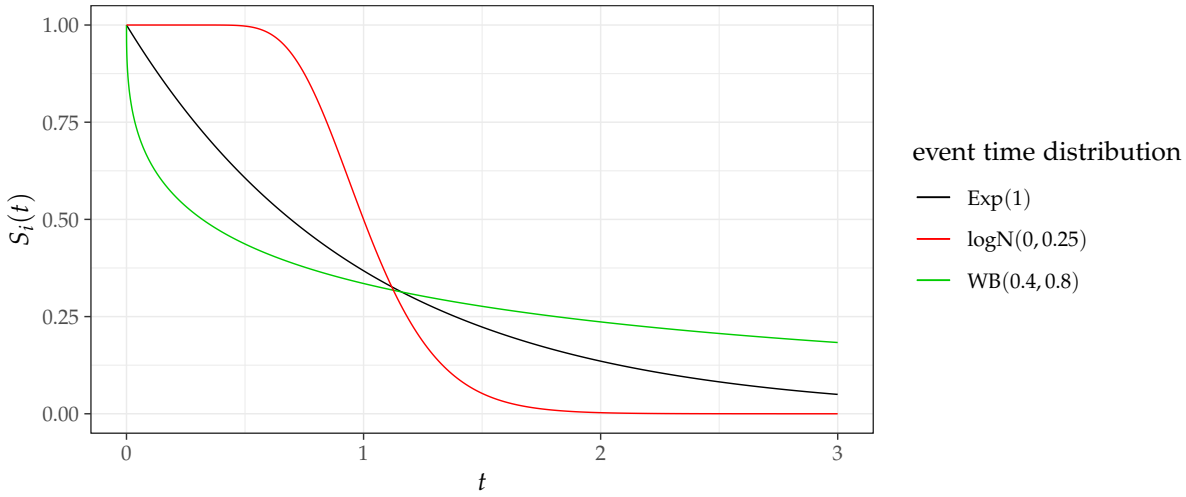

Figure 2: Event time distributions for proportional curves

To include the case of crossing survival curves under the H_0 the hazard function $\lambda(t) = 3 \cdot \mathbb{1}_{[0,\frac{1}{6}]}(t) + \frac{12}{35} \cdot \mathbb{1}_{(\frac{1}{6},\infty)}(t)$ is constructed as a piecewise constant function as in the family of piecewise exponential distribution (here denoted as pwExp) to match the exponential distribution with parameter 1. The Survival function can be derived by $S(t) = \exp(-\Lambda(t))$ using the calculation for cumulative hazard functions

$$
\begin{aligned}
\Lambda(t) &= \int_0^t \lambda(u)du \\
&= \int_0^t 3 \cdot \mathbb{1}_{[0,\frac{1}{6}]}(u)du + \int_0^t \frac{12}{35} \cdot \mathbb{1}_{(\frac{1}{6},\infty)}(u)du \\
&= \begin{cases} \int_0^t 3du + \int_0^t \frac{12}{35}du \text{ , for } t \in [0, \frac{1}{6}] \\ \int_0^{\frac{1}{6}} 3du + \int_{\frac{1}{6}}^t \frac{12}{35}du \text{ , for } t > \frac{1}{6} \end{cases} \\
&= 3t \cdot \mathbb{1}_{[0,\frac{1}{6}]}(t) + \left(\frac{3}{6} + \frac{12}{35}\left(t - \frac{1}{6}\right)\right) \cdot \mathbb{1}_{(\frac{1}{6},\infty)}(t).
\end{aligned}
$$

Crossing hazard rates are a phenomenon, that occurs in treatments with time delayed effects. Surgical interventions, for example, yield a higher risk shortly after operation, which clearly decreases in the long term. This is also well known in chemotherapy and treatment of manic depression with psycho-pharmaceuticals, regarding suicidal events. To meet the hypotheses it is important, that the RMSTs of both distributions are the same. This can be achieved by practically selecting $\tau = 1.478929$, which yields equality of RMSTs. Figure 3 illustrates both survival functions and the time endpoint τ, which ensure the conformity regarding the null hypothesis.

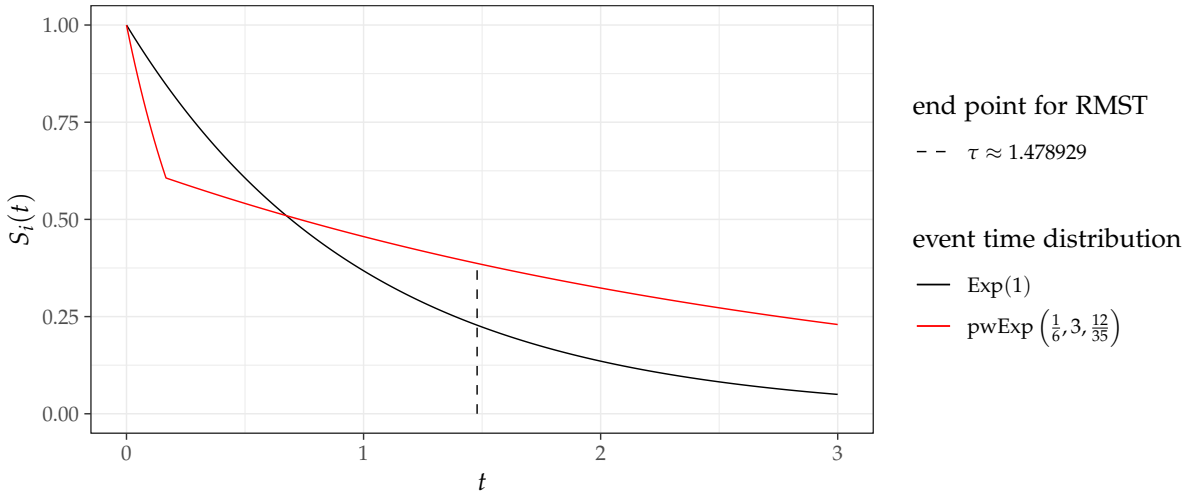

Figure 3: Crossing curve scenario under $\mathcal{H}_0$

The groups (i_A, i_B) selected to be different from Exp(1) are $(1,1),(1,2),(1,3)$ for the main effect A and $(1,1),(1,2)$ for the interaction effect. This does not imply deviation from the $\mathcal{H}_0$, as τ is chosen to lead to the same RMSTs. However appropriate considering practical data situations, this scenario is already very complex.
At last, simulations of fixed censoring rates have to be addressed. In case of $T_i \sim \text{Exp}(1)$ censoring times can be chosen to be $C_i \sim \text{Exp}\left(\frac{c_r}{1-c_r}\right)$ with censoring rate c_r. This is not valid under arbitrary event time distributional assumption. One possible solution to obtain a certain censoring rate, is to assume continuous uniform distributed censoring times $C_i \sim \text{U}(0, U_i)$ with upper bound U_i. The underlying idea is to spread the probability mass equally by scaling U_i and obtain the desired censoring rate. This incorporates a rather strong assumption. In practice, pseudo random data from the desired event time distribution is generated and U_i is determined by optimizing $P(T_i > C_i)$, replacing distributional quantities by empirical counterparts. Simplifying the expression in the following manner

$$\begin{aligned}
P(T_i > C_i) &= \int_0^\infty \min\left(\frac{x}{U_i}, 1\right) dF_i(x) \\
&= \int_0^{U_i} \frac{x}{U_i} dF_i(x) + \int_{U_i}^\infty 1 dF_i(x) \\
&= \frac{1}{U_i} \int_0^{U_i} x dF_i(x) + P(U_i \leq T_i < \infty) \\
&= \underbrace{\frac{1}{U_i} \int_0^{U_i} x dF_i(x)}_{1} + \underbrace{1 - F_i(U_i)}_{2},
\end{aligned}$$

with $F_i(x)$ as the cumulative distribution function of event times T_i, guides the practical implementation. For small censoring rates (broader distributed censoring times) the

terms 1 and 2 dominate the integral. Higher censoring times, however lead to a bigger upper limit, which enables to use the approximation $\frac{1}{U_j}\mathrm{E}(T_j)$ for the part 1 whereas part 2 becomes very small. Finally the distributional specifications can be written in the following manner:

$$
\begin{aligned}
&T_{ij} \sim \mathrm{Exp}(1),\ C_{ij} \sim \mathrm{Exp}\left(\frac{c_r}{1-c_r}\right) \text{ with } \tau = 1,\\
&T_{ij} \sim \mathrm{Exp}(1) \text{ and } T_{ij} \sim \mathrm{pwExp}\left(\frac{1}{6}, 3, \frac{12}{35}\right),\ C_{ij} \sim \mathrm{U}(0, U_i) \text{ with } \tau = 1.478929,\\
&T_{ij} \sim \mathrm{LogN}(0, 0.25),\ C_{ij} \sim \mathrm{U}(0, U_i) \text{ with } \tau = 1.5,\\
&T_{ij} \sim \mathrm{Weibull}(0.4, 0.8),\ C_{ij} \sim \mathrm{U}(0, U_i) \text{ with } \tau = 1,
\end{aligned}
$$

where time end points are chosen to be $\tau < U_i$.

3.2.1. Estimated Type 1 Error Rate

Table 4 shows the estimated type 1 error rates for the before mentioned settings, including a colour coding, whether the estimated values are covered by the corresponding binomial confidence interval. The case of exponentially distributed event times in testing of the main effect exhibits a clear difference between the asymptotic and permutation test. While the estimated type 1 error rate decreases for increasing sample size, it increases for higher censoring rates in case of asymptotic testing. The type of design, balanced or unbalanced, does only increase estimated values slightly for different sizes of sampled groups. Nearly all estimated rates of the permutation test are covered by the asymptotic binomial confidence interval. For the Log Normal distributed event times, asymptotic and permutation tests show a similar behaviour as for the exponential distribution. Weibull distributed times however exhibit different error rates. While rates still decrease with sample size, they also decrease for increasing censoring rates. This aspect could be deduced to the specific shape of the survival curve in combination with the chosen censoring times of a uniform distribution. The censoring mechanism assumes same censoring times on the complete range of the supported time interval. In general, the type 1 error rate is higher in the Weibull event time case and asymptotic test procedures, compared to all other proportional curve cases. The complex non proportionality setting with crossing survival and hazard curves clearly differs in a few aspects. The error rates are undoubtedly inflated in small finite samples as well as for the permutation test, while improving substantially for increasing sample sizes. Although still not satisfactory, this tendency of the permutation test indicates the asymptotic property of consistency with a higher convergence rate than the asymptotic Wald-Type test. The curious effect of censoring rates is even more noticeable and should be investigated detailed in further approaches. It can be concluded that the permutation approach using the Wald-Type test statistics prevents from high inflated type 1 error rates.

			c_{low}		c_{medium}		c_{high}	
Distribution	Design	q	Asymp	Perm	Asymp	Perm	Asymp	Perm
Exp	n_{bal}	1	0.1202	0.0414	0.1404	0.0466	0.1982	0.0444
		2	0.089	0.052	0.1062	0.0478	0.1444	0.049
		4	0.0672	0.0478	0.0834	0.0512	0.108	0.048
	n_{unbal}	1	0.1314	0.0484	0.1566	0.0464	0.2136	0.0492
		2	0.0898	0.0468	0.1086	0.048	0.1466	0.0474
		4	0.0704	0.0468	0.0864	0.0558	0.107	0.047
Exp-pwExp	n_{bal}	1	0.4012	0.17	0.3478	0.1282	0.2544	0.08
		2	0.2596	0.1322	0.2206	0.1024	0.1564	0.0686
		4	0.1726	0.0992	0.1426	0.0762	0.1136	0.0628
	n_{unbal}	1	0.3642	0.1248	0.3202	0.0976	0.2476	0.0634
		2	0.2442	0.1032	0.1978	0.0792	0.1516	0.0566
		4	0.1478	0.0736	0.1254	0.062	0.1034	0.0504
LogN	n_{bal}	1	0.1222	0.0538	0.1224	0.0446	0.1542	0.0446
		2	0.0786	0.0494	0.079	0.0458	0.1074	0.0552
		4	0.0632	0.0488	0.0738	0.0538	0.0792	0.0516
	n_{unbal}	1	0.121	0.047	0.1312	0.0456	0.1584	0.0474
		2	0.075	0.0424	0.0922	0.05	0.1046	0.0482
		4	0.0668	0.051	0.0712	0.053	0.0792	0.0498
Weibull	n_{bal}	1	0.3154	0.0512	0.2806	0.0468	0.2274	0.0466
		2	0.2256	0.046	0.2102	0.0482	0.1538	0.054
		4	0.1456	0.0472	0.1376	0.0488	0.1084	0.0516
	n_{unbal}	1	0.3366	0.0504	0.2966	0.0488	0.2482	0.0568
		2	0.2382	0.046	0.2104	0.0482	0.163	0.0516
		4	0.1666	0.05	0.1446	0.0444	0.1138	0.054

Table 4: estimated error type 1 for the main effect of A, with color coded values regarding the binomial confidence interval: $\widehat{E}(\phi) \geq 0.564$, $\widehat{E}(\phi) \in (0.044, 0.564)$, $\widehat{E}(\phi) \leq 0.044$

The error rates for the asymptotic test for interaction effects are even more inflated in all set ups. Additionally it could be observed, that more conservative test decisions were made in permutation.

			c_{low}		c_{medium}		c_{high}	
Distribution	Design	q	Asymp	Perm	Asymp	Perm	Asymp	Perm
Exp	n_{bal}	1	0.1576	0.04	0.196	0.0454	0.2668	0.043
		2	0.1088	0.0494	0.125	0.044	0.1788	0.0434
		4	0.0788	0.0466	0.1018	0.0534	0.1442	0.0532
	n_{unbal}	1	0.1698	0.045	0.2166	0.0474	0.3008	0.0458
		2	0.107	0.049	0.1452	0.0594	0.2	0.0556
		4	0.0888	0.0546	0.1022	0.0524	0.1394	0.056
Exp-pwExp	n_{bal}	1	0.4002	0.073	0.3754	0.0662	0.2998	0.0536
		2	0.2612	0.0924	0.2324	0.073	0.1782	0.0538
		4	0.1514	0.0676	0.141	0.064	0.1182	0.0516
	n_{unbal}	1	0.4452	0.0954	0.4136	0.0918	0.341	0.0666
		2	0.281	0.0994	0.252	0.0858	0.2042	0.0578
		4	0.164	0.0832	0.1442	0.073	0.1312	0.058
LogN	n_{bal}	1	0.1494	0.0444	0.168	0.0448	0.2206	0.0432
		2	0.1056	0.0518	0.1038	0.0478	0.1264	0.0504
		4	0.0684	0.0466	0.0786	0.047	0.0878	0.0514
	n_{unbal}	1	0.1576	0.0436	0.1868	0.0446	0.2372	0.0486
		2	0.1018	0.0446	0.107	0.0488	0.1392	0.0526
		4	0.0824	0.0528	0.0814	0.0516	0.0922	0.0508
Weibull	n_{bal}	1	0.474	0.0376	0.4398	0.0358	0.3408	0.043
		2	0.329	0.0406	0.2964	0.0404	0.2238	0.0488
		4	0.219	0.0468	0.1868	0.0498	0.1412	0.0522
	n_{unbal}	1	0.5026	0.0498	0.4652	0.0424	0.3594	0.043
		2	0.3566	0.0536	0.3282	0.056	0.2378	0.0536
		4	0.249	0.0612	0.2148	0.0576	0.1426	0.0534

Table 5: estimated error type 1 for the main effect of AB, with color coded values regarding the binomial confidence interval: $\widehat{\mathrm{E}}(\phi) \geq 0.564$, $\widehat{\mathrm{E}}(\phi) \in (0.044, 0.564)$, $\widehat{\mathrm{E}}(\phi) \leq 0.044$

In conclusion it should be stressed, that only for tests with appropriate type 1 error control a valid statistical inference can be performed. The used approach of random permutations is a possible method to achieve a satisfactory rate of wrongfully rejected null hypotheses.

3.3. Simulations under $\mathcal{H}_1$: Power

Tackling the practical most relevant problem of non-proportional and crossing hazard functions the setting used in the simulations under $\mathcal{H}_0$ is extended to evaluate the power of the presented tests. Instead of varying the time end point τ to disturb the null hypothesis, the discontinuity point s in the modified hazard rate $\lambda(t) = 3 \cdot \mathbb{1}_{[0,s]}(t) + \frac{12}{35} \cdot \mathbb{1}_{(s,\infty)}(t)$ will be changed. This seems to be an appropriate method, assuming treatments of varying time delayed (by s) effect, compared to a control group with underlying exponentially distributed event times with rate parameter 1. To increase interpretability of the estimated power functions, the discontinuity points s

are chosen in a fashion, that the difference

$$\Delta(s|\tau) = \text{RMST}_{\text{Exp}(1)}(\tau) - \text{RMST}_{\text{pwExp}(s,3,12/35)}(\tau)$$

of the RMSTs increases linearly. Using practical optimization, yields Δ, which is achieved by selecting s accordingly.

Δ	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4
s	0.167	0.198	0.233	0.273	0.32	0.377	0.449	0.55	0.719

Table 6: Difference of RMSTs Δ and the corresponding discontinuity point s

Table 6 presents the resulting values. Figure 4 illustrates the survival function for all those Δ and respectively s.

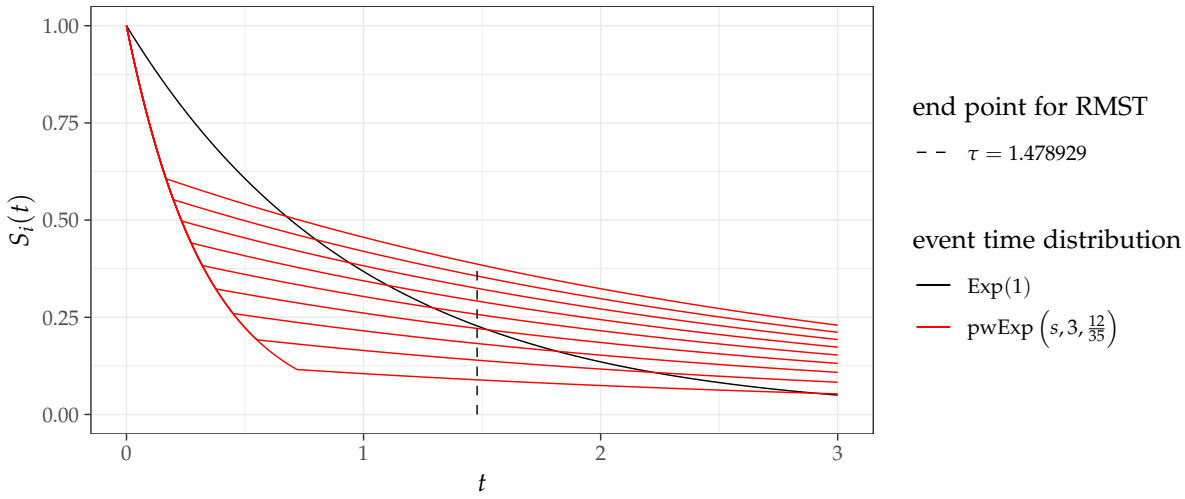

Figure 4: Crossing curves for all s

3.3.1. Estimated Power

Figure 5 presents the development of the power for the two tests (also illustrated in table 11) in relation to the predefined difference of RMSTs Δ. Starting with small samples $q = 1$ in main effect settings already highlights that, while steeply increasing power, the permutation test exhibits lower power. This is acceptable considering the hugely inflated type 1 error of the asymptotic test. For $q = 2$ and $q = 4$ the power curves draw closer to each other and lead to a very similar shape. While the type 1 error stays roughly the same for all sizes in the permutation driven test, the asymptotic test clearly improves. The power curve for medium censoring is always slightly lower than for low censoring.

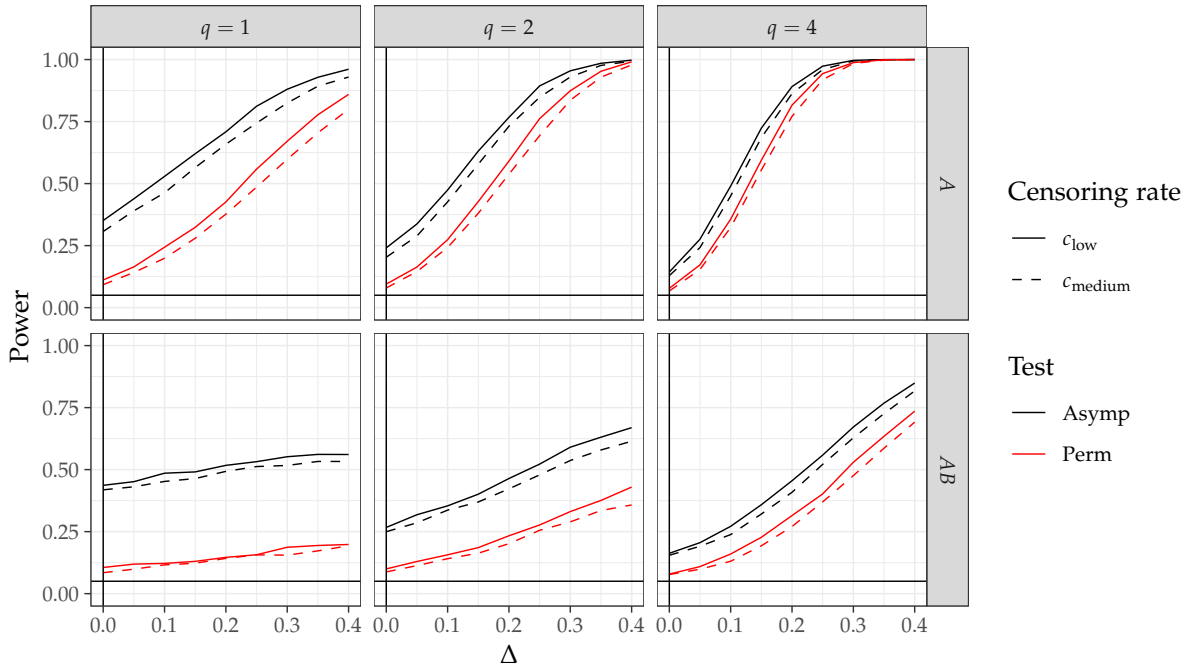

Figure 5: Power curves for crossing curve setting for differing sample sizes $q = 1, 2, 4$, hypotheses A, AB, censoring rates $c_{\text{low}}, c_{\text{medium}}$ and two tests

When examining power curves for interaction effects, it is clearly different. For small samples $q = 1$, the power curves exhibit a very flat slope with values around 0.5 (for Asymptotic) and 0.1 (for Permutation). The increase of power with higher sample sizes is quite slow and is barely acceptable for the highest sample size specification $q = 4$. Interpretation of these results should be done carefully, keeping in mind, that the situation under $\mathcal{H}_1$ was created by choosing the group (i_A, i_B) with the differing distributional assumptions. This has potentially a major influence on how easy deviations of the $\mathcal{H}_0$ can be detected. Also the trade-off between type 1 error rate and power should be handled with particular care.

4. Real Data Example

For illustration purposes the widely known data set of colon cancer by Moertel et al. (1990), available in the `R`-package `condSURV`, is used. In the controlled clinical trial, $n = 929$ patients with colon cancer with extensive metastasis and regional invaded lymph nodes (stage C) where treated, which indicates a particular poor prognosis. The patients received a adjuvant chemotherapy with two different medications. Patients were assigned to one of three treatment groups in a dynamic randomization method described by Pocock and Simon, 1975. The three study arms where a control group (`obs`), a group given levamisole (`Lev`), which was originally used treating parasitic worm infections and one group which was treated with a combination of levamisole and fluorouracil (`Lev+5FU`). Beside an extensive survey of histologic stages the variables of patient's sex (`sex`) and their treatment (`rx`) where noted and will be the focus of the following analysis. These two factors lead to a (2×3)-design already utilized to evaluate testing procedures in simulations.

	Obs	Lev	Lev+5FU	$\sum$
female	149	133	163	445
male	166	177	141	484
$\sum$	315	310	304	929

Table 7: Sample size design for both factor sex and rx

Table 7 reveals a well balanced design of sample sizes. Although the gender ratio is not exactly 1, all treatment groups have similar sizes. The time to death or censoring, which ever occurs first, was the main endpoint of the study. A time variable, which provides the time to recurrence was additionally provided, including the respective event indicator.

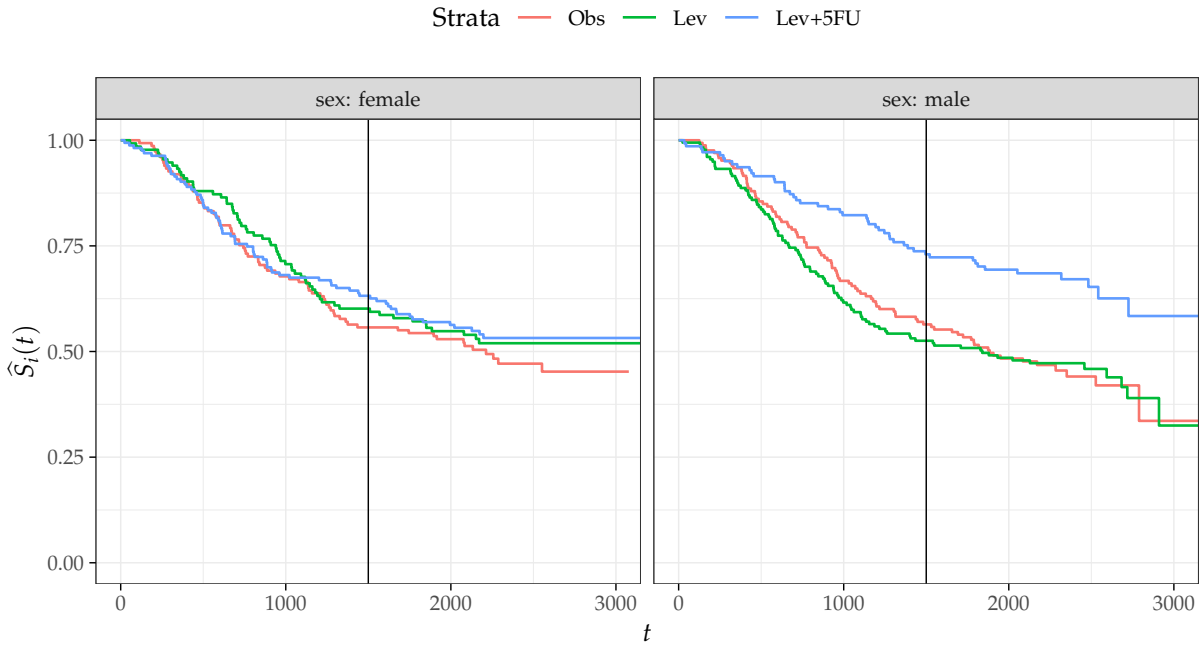

Figure 6: Survival functions for different treatments displayed for female and male

Figure 6 illustrates the Kaplan-Meier estimator for all different treatment groups side by side for both genders and already allows to gather information for possible treatment effects. It is clearly noticeable that the male group of clinical intervention with Lev+5FU shows a higher survival curve in nearly every time point observed compared to the other treatments. The other do not differ in a substantial way. All survival curves for female participants exhibit similar shapes without big differences. Comparing the genders however reveals one clearly higher survival curve for males with Lev+5FU, but also similar shapes with a steeper decrease in advanced time. In the chosen situation the median survival time is not suitable, because in two curves a value of 0.5 is not reached. This can be deduced to the censoring rates displayed in Table 8.

	Obs	Lev	Lev+5FU
female	0.48	0.53	0.54
male	0.45	0.45	0.66

Table 8: observed rate of censoring, means the proportion of observations, that had no event

Summarizing the characteristics of the estimated event time distributions leads to the RMST. The endpoint τ will now be chosen to be 1500 days, as it is roughly half of the maximal observed survival time. This choice is not necessarily clinically meaningful and is only for illustration purposes. Estimating the RMST for all subgroups shows a very similar picture in Table 9. The male group of the combination treatment exhibits a clearly higher expected survival time restricted to $\tau = 1500$.

	Obs	Lev	Lev+5FU
female	1119	1091	1144
male	1154	1074	1301

Table 9: Estimated RMSTs for all subgroups

Addressing the question if any of the treatments or the sex affects the survival time significantly the hypotheses $\mathcal{H}_0(\boldsymbol{H}_{\texttt{sex}})$, $\mathcal{H}_0(\boldsymbol{H}_{\texttt{rx}})$ and $\mathcal{H}_0(\boldsymbol{H}_{\texttt{sex,rx}})$ are formulated in the before mentioned manner. The significance level is chosen to be $\alpha = 0.05$ and to illustrate potential differences in test results the asymptotic and the permutation test are applied. The p-values displayed in Table 10 are additionally adjusted using the Bonferroni-Holm method as the test situation exhibits a multiple testing problem, which influences the type 1 error rate.

	$\mathcal{H}_0(\boldsymbol{H}_{\texttt{sex}})$	$\mathcal{H}_0(\boldsymbol{H}_{\texttt{rx}})$	$\mathcal{H}_0(\boldsymbol{H}_{\texttt{sex, rx}})$
Perm	0.106 (0.212)	0.011 (0.033)	0.141 (0.212)
Asymp	0.034 (0.057)	0 (0)	0.028 (0.057)

Table 10: unadjusted p-values for all hypotheses and both significance tests; Bonferroni-Holm adjusted p-values in parentheses

Focussing on the unadjusted p-values reveals a significant effect of the treatment (rx) detected by the permutation test. The asymptotic test in contrast rejects all three null hypotheses for $\alpha = 0.05$. Adjusting the p-values in regard to type 1 error multiplicity for every test separately leads only to the significant effect of treatment in both cases. Especially including the low power of both tests for the interaction effect, considering a more liberal interpretation of p-values could be appropriate. This seems to be even more relevant for a relatively big sample, which is reasonably balanced. A simultaneous texting procedure for multiple hypotheses could be performed by modifying the contrast matrix

5. Discussion

The approached goal of the present work was to incorporate an asymptotic exact test for the clinically meaningful effect measure of the RMST in factorial designs. The need for interpretable measures in clinical trials, especially when assumptions of classical methods do not hold, is high. Therefore the RMST was chosen to be matched to well known test procedures in factorial design. The RMST, although sensible towards the choice of time end point τ, is meaningful as expected survival time. This is easily comprehensible for medical practitioners and patients and leads to a better clinical decision making, if different treatment approaches are present. An asymptotic Wald type test statistic was used to be compared to its permutation counterpart in extensive simulations. In those, the permutation test exhibit the best properties to be applied on finite sample data, to prevent rapid inflation of the type 1 error rate. The statistical power however is clearly lower, although improving for increasing sample sizes. A considerable difference is pronounced between testing for main and interaction effects, as the latter is more difficult to reveal. This however could be a possible implication of the method and how the $\mathcal{H}_1$ was constructed. This could be an interesting topic for more detailed simulations. Additionally the determination of a fixed censoring rate is based on a strong assumption towards the censoring time distribution and limits the general applicability of the results. To overcome this limitation, additional scenarios were useful, in which more censoring time distributions would be used. The example of the colon cancer data set illustrates the difference between both significance tests. It is clearly meaningful to consider the trade-off between type 1 error rate and power behaviour. It would further be interesting to approach test methods investigating the effect of single significant factor levels, to achieve more detailed inference.

A. Additional Figures and Tables

		c_{low}		c_{medium}	
q	Δ	Asymp	Perm	Asymp	Perm
1	0	0.3512	0.1116	0.3070	0.0926
	0.05	0.4388	0.1642	0.3902	0.1418
	0.1	0.5290	0.2436	0.4628	0.1990
	0.15	0.6204	0.3236	0.5646	0.2796
	0.2	0.7088	0.4260	0.6588	0.3770
	0.25	0.8118	0.5582	0.7442	0.4858
	0.3	0.8804	0.6708	0.8250	0.5980
	0.35	0.9286	0.7776	0.8926	0.7046
	0.4	0.9610	0.8596	0.9296	0.7996
2	0	0.2410	0.0948	0.2030	0.0792
	0.05	0.3372	0.1636	0.2888	0.1454
	0.1	0.4742	0.2738	0.4264	0.2424
	0.15	0.6310	0.4274	0.5782	0.3802
	0.2	0.7672	0.5906	0.7312	0.5380
	0.25	0.8938	0.7618	0.8480	0.6932
	0.3	0.9546	0.8742	0.9298	0.8360
	0.35	0.9850	0.9528	0.9764	0.9294
	0.4	0.9976	0.9910	0.9938	0.9784
4	0	0.1436	0.0780	0.1292	0.0670
	0.05	0.2758	0.1724	0.2422	0.1538
	0.1	0.4902	0.3564	0.4502	0.3234
	0.15	0.7234	0.5940	0.6848	0.5548
	0.2	0.8912	0.8162	0.8628	0.7710
	0.25	0.9734	0.9434	0.9596	0.9190
	0.3	0.9968	0.9878	0.9944	0.9842
	0.35	0.9998	0.9990	0.9994	0.9978
	0.4	1.0000	1.0000	1.0000	0.9998

Table 11: Power values for crossing curve setting

References

A'Hern, Roger P (2016). *Restricted mean survival time: an obligatory end point for time-to-event analysis in cancer trials?*

Cox, David R (1972). "Regression models and life-tables". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 34(2), pp. 187–202.

Dahl, David B., David Scott, Charles Roosen, Arni Magnusson, and Jonathan Swinton (2019). *xtable: Export Tables to LaTeX or HTML*. R package version 1.8-4. URL: https://CRAN.R-project.org/package=xtable.

Ditzhaus, Marc, Roland Fried, and Markus Pauly (2019). "QANOVA: Quantile-based Permutation Methods For General Factorial Designs". In: *arXiv preprint arXiv:1912.09146*.

Ditzhaus, Marc, Arnold Janssen, and Markus Pauly (2020). "Permutation inference in factorial survival designs with the CASANOVA". In: *arXiv preprint arXiv:2004.10818*.

Fisher, Ronald A (1992). "The arrangement of field experiments". In: *Breakthroughs in statistics*. Springer, pp. 82–91.

Hemerik, Jesse and Jelle Goeman (2017). "Exact testing with random permutations". In: *Test* 27(4), pp. 811–825.

Henry, Lionel and Hadley Wickham (2020). *purrr: Functional Programming Tools*. R package version 0.3.4. URL: https://CRAN.R-project.org/package=purrr.

Kalbfleisch, John D and Ross L Prentice (1981). "Estimation of the average hazard ratio". In: *Biometrika* 68(1), pp. 105–112.

Kassambara, Alboukadel, Marcin Kosinski, and Przemyslaw Biecek (2020). *survminer: Drawing Survival Curves using 'ggplot2'*. R package version 0.4.7. URL: https://CRAN.R-project.org/package=survminer.

Kim, Dae Hyun, Hajime Uno, and Lee-Jen Wei (2017). "Restricted mean survival time as a measure to interpret clinical trial results". In: *JAMA cardiology* 2(11), pp. 1179–1180.

Lin, Danyu Y and Lee-Jen Wei (1989). "The robust inference for the Cox proportional hazards model". In: *Journal of the American statistical Association* 84(408), pp. 1074–1078.

Moertel, Charles G, Thomas R Fleming, John S Macdonald, Daniel G Haller, John A Laurie, Phyllis J Goodman, James S Ungerleider, William A Emerson, Douglas C Tormey, John H Glick, et al. (1990). "Levamisole and fluorouracil for adjuvant therapy of resected colon carcinoma". In: *New England Journal of Medicine* 322(6), pp. 352–358.

Montgomery, Douglas C (2013). "Montgomery Design and Analysis of Experiments Eighth Edition. Arizona State University". In: *Copyright* 2009(2005), p. 2001.

Murdoch, Duncan (2020). *tables: Formula-Driven Table Generation*. R package version 0.9.3. URL: https://CRAN.R-project.org/package=tables.

Pocock, Stuart J and Richard Simon (1975). "Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial". In: *Biometrics*, pp. 103–115.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Royston, Patrick and Mahesh KB Parmar (2013). "Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome". In: *BMC medical research methodology* 13(1), p. 152.

Terry M. Therneau and Patricia M. Grambsch (2000). *Modeling Survival Data: Extending the Cox Model*. New York: Springer. ISBN: 0-387-98784-3.

Tian, Lu, Haoda Fu, Stephen J Ruberg, Hajime Uno, and Lee-Jen Wei (2018). "Efficiency of two sample tests via the restricted mean survival time for analyzing event time observations". In: *Biometrics* 74(2), pp. 694–702.

Wickham, Hadley (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN: 978-3-319-24277-4. URL: https://ggplot2.tidyverse.org.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller (2020). *dplyr: A Grammar of Data Manipulation*. R package version 0.8.5. URL: https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley and Lionel Henry (2020). *tidyr: Tidy Messy Data*. R package version 1.1.0. URL: https://CRAN.R-project.org/package=tidyr.

Zhao, Lihui, Brian Claggett, Lu Tian, Hajime Uno, Marc A Pfeffer, Scott D Solomon, Lorenzo Trippa, and LJ Wei (2016). "On the restricted mean survival time curve in survival analysis". In: *Biometrics* 72(1), pp. 215–221.

Zhao, Lihui, Lu Tian, Hajime Uno, Scott D Solomon, Marc A Pfeffer, Jerald S Schindler, and Lee Jen Wei (2012). "Utilizing the integrated difference of two survival functions to quantify the treatment contrast for designing, monitoring, and analyzing a comparative clinical study". In: *Clinical trials* 9(5), pp. 570–577.