Inferring median survival differences in general factorial designs via permutation tests

Marc Ditzhaus¹, Dennis Dobler² and Markus Pauly¹

Abstract
Factorial survival designs with right-censored observations are commonly inferred by Cox regression and explained by means of hazard ratios. However, in case of non-proportional hazards, their interpretation can become cumbersome; especially for clinicians. We therefore offer an alternative: median survival times are used to estimate treatment and interaction effects and null hypotheses are formulated in contrasts of their population versions. Permutation-based tests and confidence regions are proposed and shown to be asymptotically valid. Their type-I error control and power behavior are investigated in extensive simulations, showing the new methods’ wide applicability. The latter is complemented by an illustrative data analysis.

Keywords
Censoring, interaction, median, resampling, survival

1 Introduction
Medical studies with time-to-event endpoints are most commonly inferred by means of logrank tests or Cox regression models. In case of non-proportional hazards, however, they are not always the most appropriate choice and several alternatives to logrank tests have been proposed.¹⁻⁶ In factorial survival designs, this issue is additionally hampered by the desire to summarize main treatment and interaction effects in single quantities. For hazard ratios as effect estimates, this can only be achieved under the proportional hazards assumption. As this can be hard to verify or simply wrong, some alternatives to hazard ratios as effect sizes are desired that provide “clinically meaningful outcomes for (…) patients” and for “discussing the value of clinical trials with patients”.⁷ Proposed alternatives cover average hazard ratios,⁸,⁹ concordance and Mann–Whitney effects,¹⁰⁻¹² the (restricted) mean survival,¹³,¹⁴ or median survival times (MSTs).¹⁵⁻¹⁷

Except for Mann–Whitney effects and concordance odds,¹¹,¹⁸ statistical tools for inferring these quantities have mostly been developed for special two or multiple sample settings. Applying these methods to factorial designs with at least two crossed factors, would neither use the factorial structure to full capacity nor allow for the quantification of potential interaction effects. But this is one of the key reasons to plan studies with a factorial design.

The present paper aims to increase the urgently needed flexibility to apply appropriate procedures that take full advantage of factorial structures in the analysis of time-to-event data.¹⁹ As the handling of the above-mentioned effect sizes requires different approaches of diverse complexity, it is impossible to treat all of them simultaneously. We thus focus on the MSTs as effect estimates and leave factorial extensions of the others for future research.

¹Faculty of Statistics, TU Dortmund University, Dortmund, Germany
²Department of Mathematics, Vrije Universiteit Amsterdam, Amsterdam, Netherlands

Corresponding author:
Marc Ditzhaus, Faculty of Statistics, Mathematical Statistics and Applications in Industry, TU Dortmund University, Vogelpothsweg 87, 44227 Dortmund, Germany.
Email: marc.ditzhaus@tu-dortmund.de
MSTs are “easy to understand” and therefore frequently reported in medical studies. In fact, they are “the most common measure used in the outcome reporting of oncology clinical trials” and their application has recently been propagated. Besides their use as a descriptive tool, they are also applied for inferential conclusions. The methodological foundation for the latter was laid in the works of Brookmeyer and Crowley and later extended by others. All these investigations focus on specific aspects of two or k sample settings and do not cover potential interactions. However, for non-censored observations, extensions to general factorial designs have recently been proposed. These are based on the idea of using heterogeneity robust studentized permutation tests in statistics of Wald-type. Thus, it appeared natural to transfer these ideas to time-to-event situations. It thereby turned out that censoring makes theoretical investigations much more demanding. Moreover, the assumption of a general censoring mechanism, that can be arbitrary across groups, additionally hampers the construction of valid estimators. All this made the study of the robust permutation approach even more difficult. Carefully extending some results on empirical processes, we nevertheless accomplished this which finally resulted in asymptotically valid and consistent studentized permutation tests for general factorial survival designs. These tests are even finitely exact in the special case of exchangeable data.

The paper is organized as follows. First, the model and all mathematical notations are introduced. After that, we construct a sensible test statistic and analyze its asymptotic properties. Subsequently, its permutation version is introduced and analyzed. To empirically assess the performance of the test, we conducted extensive simulation studies in which also a competitor from quantile regression is included for comparisons. Furthermore, a real data analysis focuses on data from trials on liver diseases. We conclude the paper with a discussion section. All proofs are given in the supplement.

2 Factorial survival set-up

To simplify notation, we regard any factorial design for now as a k-sample set-up

\[ T_{ij} \sim S_i, \quad C_{ij} \sim G_i, \quad i = 1, \ldots, k, \quad j = 1, \ldots, n_i \]

where \( T_{ij} \) and \( C_{ij} \) respectively denote the independent survival and censoring times of individual \( j \) in group \( i \). The survival functions \( S_i \) are assumed to be continuous with density functions \( f_i \); the censoring functions \( G_i \) need not be continuous. The actually observable data consist of the right-censored event times \( X_{ij} = \min(T_{ij}, C_{ij}) \) and the corresponding censoring statuses \( \delta_{ij} = 1 \{ X_{ij} = T_{ij} \} \). In this paper, the quantities of main interest are the group-specific (survival) medians

\[ m_i = S_i^{-1}(1/2) = \inf\{ t \in \mathbb{R} : S_i(t) \leq 1/2 \}, \quad i = 1, \ldots, k \]

We wish to test for relationships of the medians in various factorial designs

\[ H_0(H) : Hm = 0 \]

where \( m = (m_1, \ldots, m_k)' \in \mathbb{R}^k \), \( 0 \in \mathbb{R}^l \) is the zero vector, and \( H \in \mathbb{R}^{k \times k} \) denotes a contrast matrix, i.e. the rows sum up to 0. For example, the centering matrix \( H = P_k = I_k - J_k/k \) is used to describe the null hypothesis \( H_0(P_k) : m_1 = \ldots = m_k \) of no group effects in the k-sample setting, where \( I_k \) is the \( k \times k \)-dimensional unit matrix and \( J_k \in \mathbb{R}^{k \times k} \) consists of 1’s only.

Two-way designs are incorporated by splitting up the indices \( i = (i_A, i_B) \) for factors A (with \( a \in \mathbb{N} \) levels) and B (with \( b \in \mathbb{N} \) levels). Then, the contrast matrices for testing hypotheses about no main or interaction effects are given by

- \( H_A = P_a \otimes (J_b/b) \) with \( H_0(H_A) = \{ \bar{m}_1 = \ldots = \bar{m}_a \} \) (no main effect of factor A),
- \( H_B = (J_a/a) \otimes P_b \) with \( H_0(H_B) = \{ \bar{m}_1 = \ldots = \bar{m}_b \} \) (no main effect of factor B),
- \( H_{AB} = P_a \otimes P_b, \quad H_0(H_{AB}) = \{ \bar{m}_{i_1}i_2 - \bar{m}_{i_1} - \bar{m}_{i_2} + \bar{m} = 0 \text{ for all } i_1, i_2 \} \) (no interaction).

Here, \( \otimes \) denotes the Kronecker product and \( \bar{m}_{i_1}, \bar{m}_{i_2}, \) and \( \bar{m} \) are the means over the dotted indices. Extensions to higher-way crossed or hierarchically nested layouts can be obtained similarly.
Estimates of the survival medians will be based on the Kaplan–Meier estimators given by

$$
\hat{S}_i(t) = \prod_{j : X_{ij} \leq t} \left(1 - \frac{\delta_{ij}}{Y_i(X_{ij})}\right), \quad t \geq 0, \quad i = 1, \ldots, k
$$

Here, $Y_i(t) = \sum_{j=1}^{n_i} 1\{X_{ij} \geq t\}$ is the number of subjects in group $i$ being at risk just before $t$. Then the natural estimator of the median $m_i$ is

$$
\hat{m}_i = \hat{S}_i^{-1}(1/2) = \inf\{t \in \mathbb{R} : \hat{S}_i(t) \leq 1/2\}, \quad i = 1, \ldots, k
$$

In the same way, $\hat{S}_i^{-1}(q)$ is an estimator of the $q$-th quantile $S_i^{-1}(q), q \in (0, 1)$.

### 3 Asymptotic Wald-type tests

Inference on the population medians will be achieved by means of Wald-type tests. To this end, let us first introduce the so-called projection matrix\(^{25-27}\) $T = H (H H)^\dagger H$, where $(\cdot)^\dagger$ denotes the Moore–Penrose inverse. The benefits of $T$ are its symmetry and idempotence. The null hypotheses are unaffected by this change of contrast matrices;\(^{28}\) $T \mathbf{m} = 0$ holds if and only if $H \mathbf{m} = 0$. Then a Wald-type test statistic for this null hypothesis is given as

$$
W_n(T) = n(T \mathbf{m})' (T \Sigma T)^\dagger T \mathbf{m} \quad \text{with} \quad \Sigma = \text{diag}\left(\frac{n}{n_1} \hat{\sigma}_1^2, \ldots, \frac{n}{n_k} \hat{\sigma}_k^2\right).
$$

Here, $n = n_1 + \ldots + n_k$ is the total sample size and $\bar{m} = (m_1, \ldots, m_k)'$ is the vector of pooled median estimators. Moreover, $\hat{\sigma}_i^2$ denote consistent estimators of the asymptotic variances $\sigma_i^2 = \sqrt{n_i} (\hat{m}_i - m_i)$ which are yet to be determined. In fact, discussing the weak convergence of $W_n(T)$ boils down to investigating the convergence of the median and subsequent variance estimators. To this end, denote by $\overset{d}{\rightarrow}$ convergence in distribution and by $\overset{p}{\rightarrow}$ convergence in probability. Throughout the paper, we make the following assumption which ensures the existence of $\hat{m}_i = \hat{S}_i^{-1}(1/2)$ with a probability tending to 1 and the asymptotic normality of $\hat{m}_i$ as $n_i \rightarrow \infty$.

**Assumption 1.** For each sample group $i = 1, \ldots, k$ it holds that

(a) The density is positive at the median: $f_i(m_i) > 0$.

(b) It is possible to observe events after the MST: $G_i(m_i) > 0$.

A variant of the following proposition traces back to Sander.\(^{28}\)

**Proposition 1.** Suppose Assumption 1. As $n_i \rightarrow \infty$, $\sqrt{n_i} (\hat{m}_i - m_i) \overset{d}{\rightarrow} Z_i \sim \mathcal{N}(0, \sigma_i^2)$ with variance

$$
\hat{\sigma}_i^2 = -\frac{1}{4 f_i(m_i)^2} \int_0^{m_i} \frac{dS_i}{G_i \hat{S}_i^2}
$$

Proposition 1 shows that estimators of the asymptotic variances should involve an estimate of $f_i(m_i)$. For computational reasons, we propose to use an estimator of the standard deviation which, in the uncensored case, is based on a standardized confidence interval;\(^{29}\) see Price and Bonett\(^{30}\) for a slight modification to improve the estimator’s performance for small sample sizes. Chakraborti\(^{31,32}\) extended the estimator to the censored case and proved its consistency. For our purposes, we use different variants of this estimator. The first is given by

$$
\hat{\sigma}_{i, \text{two-sided}} = 0.5 z_{\gamma/2} \sqrt{n_i (\hat{S}_i^{-1}(l_i) - \hat{S}_i^{-1}(u_i))}
$$

where $\gamma \in (0, 1)$ is a fixed number, $z_x = \Phi^{-1}(1-x), x \in (0, 1)$ is the $(1-x)$-quantile of the standard normal distribution function $\Phi$ and we set

$$
l_i = 0 \lor 0.5 \left(1 - z_{\gamma/2} \sqrt{\hat{V}_i/n_i}\right) \quad \text{and} \quad u_i = 1 \land 0.5 \left(1 + z_{\gamma/2} \sqrt{\hat{V}_i/n_i}\right) \quad \text{for} \quad \hat{V}_i = n_i \sum_{j : X_{ij} \leq m_i} \frac{\delta_{ij}}{\gamma(X_{ij})}
$$
Here, \( c \lor d \) and \( c \land d \) denote the maximum and minimum of two real numbers \( c \) and \( d \), respectively. We note that \( \hat{V}_i \) is a consistent estimator of the asymptotic variance of the normalized Nelson–Aalen estimator at \( m_i \).

In cases of strong censoring in combination with small sample sizes, \( S_i^{-1}(l_i) \) may not exist. Here, one solution is to adjust the involved standard normal quantile; see also Price and Bonett for a short discussion on the choice of \( \gamma \). Usually, \( \gamma = 5\% \) or \( 10\% \) is chosen. If \( S_i^{-1}(l_i) \) does not exist for the previously made choice of \( \gamma \), we suggest to use the Kaplan–Meier estimator at the last observed event time \( \hat{l}_i := \min_{t > 0} \hat{S}_i(t) \) instead of \( l_i \). In this case, we solve for \( \gamma \) in the original definition of \( l_i \), i.e.,

\[
\hat{\gamma} := 2 \Phi \left( \frac{1 - 2 \hat{l}_i}{\sqrt{\hat{V}_i/n_i}} \right)
\]

This \( \hat{\gamma} \) then gives us \( \hat{u}_i = 1 \lor 0.5(1 + z_{\hat{\gamma}/2}(\hat{V}_i/n_i)^{1/2}) \) instead of \( u_i \). Finally, the adjusted estimator of the standard deviation is obtained according to formula (6) with \( \hat{\gamma} \), \( \hat{l}_i \), and \( \hat{u}_i \) instead of \( \gamma \), \( l_i \), and \( u_i \), respectively. Note that this adjustment does not play a role as \( n_i \to \infty \) because \( \hat{S}_i^{-1}(l_i) \) exists with a probability tending to 1 for any choice of \( \gamma \).

A different approach to overcome the problem of the non-existence of \( \hat{S}_i^{-1}(l_i) \) is to switch from two-sided intervals to one-sided intervals. In this case, the estimator is given by

\[
\hat{\sigma}_{i, \text{one–sided}} = z_{\hat{\gamma}/2} \sqrt{n_i (m_i - \hat{S}_i^{-1}(u_i))}
\]

To unify notation, we subsequently suppose that \( \hat{\sigma}_i \) is either the adjusted two-sided interval variance estimator from equation (6) or the one-sided counterpart from equation (7). While the theory developed below is valid for both choices, the simulations below show a slightly advantageous power behavior for the one-sided version.

**Lemma 1.** As \( n_i \to \infty \), we have \( \hat{\sigma}_i \overset{p}{\to} \sigma_i \) for each \( i = 1, \ldots, k \).

By combining all of the convergences discussed above, we wish to find the limit null distribution of \( W_n(T) \). However, as the sample sizes might grow at different paces, we need to make the following weak assumption that, asymptotically, no sample group vanishes in relation to any other group:

**Assumption 2.** \( \min_{i = 1, \ldots, k} \liminf_{n \to \infty} n_i/n > 0 \).

Under this assumption, it follows that the vector \( \hat{m} \) of estimated medians converges in distribution at \( \sqrt{n} \)-rate.

In particular, we obtain the following theorem:

**Theorem 1.** Suppose that Assumptions 1 and 2 hold, then we have

(a) under \( \mathcal{H}_0(T) : T_m = 0 \), \( W_n(T) \overset{d}{\to} Z \sim \chi^2_{\text{rank}(T)} \) as \( n \to \infty \).

(b) under \( \mathcal{H}_1(T) \) : \( T_m \neq 0 \), \( W_n(T) \overset{d}{\to} \chi^2_{\text{rank}(T) + z} \) as \( n \to \infty \).

For \( \alpha \in (0, 1) \) denote by \( \chi^2_{\alpha} \) the \((1 - \alpha)\)-quantile of a \( \chi^2 \)-distribution with \( \ell \) in \( N \) degrees of freedom.

**Corollary 1.** Let \( \alpha \in (0, 1) \) and suppose that Assumptions 1 and 2 hold. The asymptotic test \( \varphi_n = 1\{W_n(T) > \chi^2_{\text{rank}(T) + z} \} \) for \( \mathcal{H}_0(T) \) versus \( \mathcal{H}_1(T) \) is consistent with asymptotic level \( \alpha \), i.e. as \( n \to \infty \),

\[
E(\varphi_n) = P(W_n(T) > \chi^2_{\text{rank}(T) + z}) \to 1_{\mathcal{H}_1(T)} + \alpha \cdot 1_{\mathcal{H}_0(T)}
\]

### 3.1 General quantiles

The aforementioned results and the subsequent statements about the respective permutation statistic can be directly transferred to general \( p \)-quantiles, \( p \in (0, 1) \). For the reader’s convenience, we give here the specific form of the Wald-type statistic to infer, more generally, the null hypothesis \( \mathcal{H}_0 : H(q_1, \ldots, q_k) = 0 \), \( q_i = S_i^{-1}(p) \), based on \( p \)-quantiles. First, we clearly need to replace \( \hat{m} \) in equation (4) by the vector \( q = (\hat{q}_1, \ldots, \hat{q}_k) \) of the respective empirical quantiles \( \hat{q}_i = S_i^{-1}(p) \). Second, we require an adoption of the (groupwise) variance estimators \( \hat{\sigma}_i^2 \). A direct extension of the proofs yields the following general form of the two-sided variant, see equation (6)

\[
\hat{\sigma}_{i, \text{two–sided}} = 0.5 z_{\hat{\gamma}/2} \sqrt{n_i (\hat{S}_i^{-1}(l_{i,p}) - \hat{S}_i^{-1}(u_{i,p}))}
\]
where

\[ l_{i,p} = 0 \lor p \left( 1 - z_{\gamma/2} \sqrt{\hat{V}_{i,p}/n_i} \right), \quad u_{i,p} = 1 \land p \left( 1 + z_{\gamma/2} \sqrt{\hat{V}_{i,p}/n_i} \right), \quad \hat{V}_{i,p} = \sum_{j \neq i \in \bar{q}_i} \frac{n_j \delta_{ij}}{\hat{Y}_j^2(X_i^q)} \]

In the same spirit, we obtain the updated one-sided variant \( \hat{\sigma}_{i, \text{one-sided}, p} = z_{\gamma/2} \sqrt{\bar{q}_i - \hat{S}_i^{-1}(u_{i,p})} \).

### 4 Permutation tests

Theorem 1 is already sufficient to deduce consistent tests for \( \mathcal{H}_0(\mathbf{T}) \) versus \( \mathcal{H}_1(\mathbf{T}) \). However, if the sample sizes \( n_1, \ldots, n_k \) are not large enough, the type-I error rates of the tests based on Wald-type statistics are often inflated;\(^{22,25} \) see also the simulation section below. To solve this problem, we suggest to conduct the new Wald-type test as a robust permutation test. Denote by \( \pi = (\pi(1), \pi(2), \ldots, \pi(n)) \) a random vector that is uniformly distributed on the set of all permutations of \( (1, \ldots, n) \). Writing the pooled data as

\[ \mathbf{X} = (X_1, \delta_1, X_2, \delta_2, \ldots, X_n, \delta_n) = (X_{11}, \delta_{11}, X_{12}, \delta_{12}, \ldots, X_{kn}, \delta_{kn}) \]

the permuted new groups are given as

\[ \mathbf{X}_{1, \pi}^\pi = (X_{\pi(1)}, \delta_{\pi(1)}), \ldots, \mathbf{X}_{k, \pi} = (X_{\pi(n-n_k+1)}, \delta_{\pi(n-n_k+1)}, \ldots, X_{\pi(n)}, \delta_{\pi(n)}) \]

We write \( \hat{S}_i^\pi \) and \( \hat{\sigma}_i^\pi \) for the Kaplan–Meier and standard deviation estimators from the previous section but based on \( \mathbf{X}_{i, \pi}^\pi, i = 1, \ldots, k \), respectively. Similarly, each \( \hat{m}_i^\pi \) is defined in terms of \( \hat{S}_i^\pi \). Finally, the permutation Wald-type statistic results as

\[ W_n^\pi(\mathbf{T}) = n(\mathbf{T} \hat{\sigma}^\pi)^\prime (\mathbf{T} \Sigma^\pi)^\prime \mathbf{m}^\pi \]

where \( \mathbf{m}^\pi = (\hat{m}_1^\pi, \ldots, \hat{m}_k^\pi) \) and \( \Sigma^\pi = \text{diag}(\hat{\sigma}_1^\pi, \ldots, \hat{\sigma}_k^\pi) \). Theoretical analyses from the appendix show that the permutation medians \( \hat{m}_i^\pi \) behave similarly to the pooled median estimator \( \hat{m} = \hat{S}^{-1}(0.5) \) where \( \hat{S} \) denotes the pooled Kaplan–Meier estimator. That is, \( \hat{S} \) uses the complete dataset \( \mathbf{X} \). Denote by \( S \) the limit in probability of \( \hat{S} \); the convergence is argued in the supplement where also an explicit formula for \( \hat{S} \) is given. For the existence and the convergence of the pooled median estimator and the permutation medians, we make the following assumption on the censoring distributions:

**Assumption 3**

(a) The function \( f_0 := \sum_{i=1}^k f_i \) is positive at \( m = S^{-1}(0.5) \).
(b) The pooled median is observable, i.e. \( \min_{i=1, \ldots, k} \mathbb{E}[G_i(m)] > 0 \).
(c) The functions \( f_i \) and \( G_i, i = 1, \ldots, k \), are continuous on a neighborhood of \( m \).

Under these assumptions, the studentized permutation approach works as stated below.

**Theorem 2.** Suppose that **Assumptions 1–3** hold. Under \( \mathcal{H}_0(\mathbf{T}) : \mathbf{m} = \mathbf{0} \) and also under \( \mathcal{H}_1(\mathbf{T}) : \mathbf{m} \neq \mathbf{0} \), the conditional distribution of the permutation version \( W_n^\pi(\mathbf{T}) \) of \( W_n(\mathbf{T}) \) given \( \mathbf{X} \) always approaches the null distribution of \( W_n(\mathbf{T}) \), i.e. as \( n \to \infty \)

\[ \sup_{x \in \mathbb{R}} |\Pr(W_n^\pi(\mathbf{T}) \leq x \mid \mathbf{X}) - \Pr_{\mathcal{H}_0(\mathbf{T})}(W_n(\mathbf{T}) \leq x)| \to 0 \]

This result enables the construction of a consistent permutation test for \( \mathcal{H}_0(\mathbf{T}) : \mathbf{m} = \mathbf{0} \) versus \( \mathcal{H}_1(\mathbf{T}) : \mathbf{m} \neq \mathbf{0} \): denote by \( c_n^\pi \) the \( (1 - \alpha) \)-quantile of the conditional distribution of \( W_n^\pi(\mathbf{T}) \) given \( \mathbf{X} \). Then Lemma 1 and Theorem 7 in Janssen and Pauls\(^3^4 \) ensure the convergence of \( c_n^\pi \) to \( \chi_{\text{rank}(\mathbf{T})}^2 \) in probability. This yields

**Corollary 2.** Let \( \alpha \in (0, 1) \) and suppose that **Assumptions 1–3** hold. The permutation test \( \varphi_n^\pi = 1\{W_n(\mathbf{T}) > c_n^\pi\} \) for \( \mathcal{H}_0(\mathbf{T}) \) versus \( \mathcal{H}_1(\mathbf{T}) \) is consistent and has asymptotic level \( \alpha \), i.e. as \( n \to \infty \)

\[ E(\varphi_n^\pi) = P(W_n(\mathbf{T}) > c_n^\pi) \to 1_{\mathcal{H}_1(\mathbf{T})} + \alpha \cdot 1_{\mathcal{H}_0(\mathbf{T})} \]
Beyond this asymptotic correctness for general models, the permutation test has yet another beneficial property: in the special case that all sample groups share the same underlying distributions, i.e. if the observed data pairs are exchangeable, a randomized version of the test can be shown to be finitely exact, for all possible sample size combinations. We refer to Hemerik and Goeman \(^3^4\) and the references cited therein for further reading on the exactness of permutation tests.

### 4.1 Confidence intervals

While permutation methods are famous for statistical testing, it is less well known that they can also be applied for constructing confidence intervals\(^1^0\); we also refer to the discussion by Pauly et al.\(^3^5\). Let \(\mathbf{c} \) be a contrast vector of interest, then \(\sqrt{n}\mathbf{c}'(\mathbf{m} - \mathbf{m})/\sqrt{\mathbf{c}'\mathbf{\Sigma}\mathbf{c}}\) and, conditionally on the data, \(\sqrt{n}\mathbf{c}'\mathbf{m}^\circ/\sqrt{\mathbf{c}'\mathbf{\Sigma}_{\text{c}}\mathbf{c}}\) are asymptotically standard normally distributed. Hence, permutation-based confidence intervals for \(\mathbf{c}'\mathbf{m}\) with asymptotic confidence level \((1 - \alpha)\) are for instance of the form

\[
[\mathbf{c}'\mathbf{m} - \mathbf{c}'\hat{\mathbf{\Sigma}}_{\alpha, \alpha} \sqrt{\mathbf{c}'\mathbf{\Sigma}\mathbf{c}}, \mathbf{c}'\mathbf{m} + \mathbf{c}'\hat{\mathbf{\Sigma}}_{\alpha, \alpha} \sqrt{\mathbf{c}'\mathbf{\Sigma}\mathbf{c}}]
\]

where \(\mathbf{c}'\hat{\mathbf{\Sigma}}_{\alpha, \alpha}\) denotes the \((1 - \alpha)\)-quantile of the conditional distribution of \(\mathbf{c}'\mathbf{m}^\circ/\sqrt{\mathbf{c}'\mathbf{\Sigma}_{\text{c}}\mathbf{c}}\) given the data. Recall from the previous section that all results remain valid when considering general \(p\)-quantiles instead of medians. This naturally also applies to the confidence interval described here.

Typically, permutation tests such as \(\phi_n\) and permutation-based confidence intervals respectively show a good control of the nominal type-I error probability and coverage rates close to the nominal level, even for non-exchangeable situations and in the case of small samples. We will empirically assess the finite sample behavior of the above-developed permutation tests in the following section.

### 5 Simulations

To complement our asymptotic results, we conducted an extensive simulation study to cover various small sample size settings. For ease of presentation, we focused on the following two-factorial designs.

(a) A \(2 \times 2\) layout with factors \(A\) and \(B\) (each with \(a = b = 2\) levels) in which we test for the presence of a main effect of factor \(A\) and of an interaction effect by choosing the contrast matrices \(\mathbf{H} = \mathbf{H}_A\) and \(\mathbf{H} = \mathbf{H}_{AB}\), respectively. Here, we considered two different sample size scenarios \(\mathbf{n}_1 = (n_{11}, n_{12}, n_{21}, n_{22}) = (12, 12, 12, 12)\) (balanced) and \(\mathbf{n}_2 = (16, 11, 7, 14)\) (unbalanced), and three different censoring settings with censoring rate vectors \(\mathbf{c}_1 = (c_{r11}, c_{r12}, c_{r21}, c_{r22}) = (0.07, 0.12, 0.12, 0.07)\) (low censoring), \(\mathbf{c}_2 = (0.29, 0.38, 0.25, 0.35)\) (high censoring), and \(\mathbf{c}_3 = (0.12, 0.38, 0.07, 0.29)\) (low/high censoring).

(b) A \(2 \times 3\) layout with factors \(A\) (possessing \(a = 2\) levels) and \(B\) (with \(b = 3\) levels) in which we wish to test the null hypotheses of no main effect of \(A\) by considering the matrix \(\mathbf{H} = \mathbf{H}_A\). For this scenario, we chose two sample size settings \(\mathbf{n}_1 = (n_{11}, n_{12}, \ldots, n_{23}) = (10, \ldots, 10)\) (balanced) and \(\mathbf{n}_2 = (8, 10, 12, 8, 10, 12)\) (unbalanced), as well as three different censoring scenarios given by the censoring rate vectors \(\mathbf{c}_4 = (0.15, 0.2, 0.25, 0.15, 0.2, 0.25)\) (medium censoring I), \(\mathbf{c}_5 = (0.3, 0.25, 0.2, 0.3, 0.25, 0.2)\) (medium censoring II), and \(\mathbf{c}_6 = (0.05, 0.1, 0.15, 0.2, 0.15, 0.1)\) (low censoring).

#### 5.1 Distributional choices

The group-wise survival times \(T_i, i = (i_A, i_B)\), were simulated according to one of the following distributions:

1. a standard exponential distribution \((\text{Exp})\),
2. a Weibull distribution with parameters \(\lambda_{\text{shape}} = 2\) and \(\lambda_{\text{scale}} = \log(2)^{-1/2}\) \((\text{Weib})\),
3. a standard log-normal distribution \((\text{LogN})\),
4. different mixture settings \((\text{mix})\).

The first three distributions were used for the first three sample groups in both factorial designs (a) and (b). In the \(2 \times 2\) layout (a), the fourth group was generated according to a mixture of all three other distributions. However, in the \(2 \times 3\) layout (b), the fourth, fifth, and sixth groups are pairwise mixtures of the distributions 1 and 3, 1 and 2, 2 and 3, respectively. For the censoring times, we considered uniform distributions \(\text{Unif}[0, U]\)
where the endpoint $U_i$ in group $i$ was calibrated by a Monte-Carlo simulation such that the average censoring rate equals the pre-chosen $cr_i$. For this, the formula $cr_i = P(T_i > C_i) = -\int_0^{\infty} \min\{x/U_i, 1\} \ dS_i(x)$ was used. The settings described above correspond to the null hypotheses and were used to study the tests’ type-1 errors in the first subsection that follows below.

To obtain respective alternatives, we shifted the survival and censoring distributions of the first group in the $2 \times 2$ layout and of the first two groups for the $2 \times 3$ layout by $\delta \in \{0.2, 0.4, 0.6, 0.8, 1\}$ to the right. The tests’ power performances for these shift alternatives are analyzed in the final simulation subsection.

### 5.2 Competing methods

We included the asymptotic Wald-type tests using the one- and two-sided interval variance estimators with $\gamma = 10\%$ as well as the respective permutation counterparts. Due to the partially high censoring settings, we chose the adaptive procedure introduced in the above section “Asymptotic Wald-type tests” for the two-sided interval strategy. For example, in the scenario with exponentially distributed survival times, $n = n_4$ and $cr = cr_4$, the upper confidence interval limit $\hat{S}_i(l_i)$ did not exist for at least one of the groups in 42% of all considered iterations. Thus, the adaptive procedure with $\hat{\gamma}$, $\hat{l}_i$, and $\hat{u}_i$, instead of $\gamma$, $l_i$, and $u_i$, which we use here, reduces the number of such failed iteration runs, while it coincides with the original two-sided interval method whenever $\hat{S}_i(l_i)$ can be determined. We compared these four different testing procedures with a quantile regression method for survival data applied to the median.\(^36\) The evaluation of the latter was done by means of the jackknife method,\(^37\) which is the default choice in the R-package `quantreg`.\(^38\) Within the regression, the factors were treated as nominal variables and represent, for example, different treatments or different (ethnic) origins. In addition, the interactions of the factors were included to get a fair comparison with the newly developed Wald-type tests.

The simulations were conducted by means of the computing environment R,\(^39\) version 3.6.2, generating $N_{\text{sim}} = 5000$ simulation runs and $N_{\text{res}} = 1999$ resampling iterations for the permutation and jackknife procedures. In the rare case that any sample group-specific median does not exist, the corresponding simulation run was not included in the analysis. This mostly happened (in around 6% of the cases) for the set-up with censoring $cr = cr_3$, $n = n_2$ and log-normally or exponentially distributed survival times. The nominal significance level was set to $\alpha = 5\%$.

### 5.3 Type-1 error

In this subsection, we analyze the type-1 error rate behavior of all five tests. For an assessment of the results, we recall that the standard error of the estimated sizes for the $N = 5000$ simulation runs is 0.3% if the true type-1 error probability is indeed 5%. This would yield the binomial confidence interval $[4.4\%, 5.6\%]$ for the estimated sizes. Values outside this interval deviate significantly from the nominal 5% level.

The simulation results for the $2 \times 2$ design are presented in Table 1. It is readily seen that the jackknife approach for the quantile regression method keeps the nominal level quite accurately with a slight tendency to liberal and conservative decisions for tests of no main and no interaction effect, respectively. The two permutation Wald-type tests control the nominal level reasonably well, except for the scenarios with simultaneous low and high censoring. Here, the two-sided interval variance estimation leads to rather liberal decisions with values up to 7.6%, while the one-sided approach only exhibits a slight liberality under the mixture distribution setting. In contrast, the type-1 error probabilities of the asymptotic Wald-type tests strongly depend on the chosen variance estimator: on the one hand, the one-sided approach exhibits rather liberal results, which are most pronounced in the low-censoring settings with rejection rates up to 13.8%. An exception is the high censoring setting under lognormal distributions; here, the decisions are rather conservative with values between 3.6% and 4.4% for the small sample sizes. On the other hand, the two-sided strategy leads to rather conservative decisions with values down to 3.2%, but the values are frequently (41 of 96 settings) contained in the binomial confidence interval.

The type-1 error rates for the $2 \times 3$ design are displayed in Table 2. In contrast to the previous observations, the jackknife approach now behaves very liberally with values between 8.7% and 11.9%. Both permutation tests keep the type-1 error satisfactorily: under the mixture distribution settings, we can find a slight tendency to liberal decisions with values up to 6.6% and 6.1% for the two- and one-sided variance estimation strategy, respectively. While the asymptotic Wald-type test with the two-sided approach frequently keeps the type-1 error in the $2 \times 2$ layout, the decisions are now quite conservative. The conservative behavior is most pronounced in the medium censoring II setting for the lognormal distributions with values as small as 1.9%. Comparing this to the one-sided
strategy, the decisions again become rather liberal with values up to 10.9% in the medium censoring II setting. However, they come closer to the 5% benchmark when the sample sizes are doubled. An exception to this liberal behavior can be observed under the Weibull distribution, where the decisions are rather conservative with values between 3.6% and 5.5%.

In summary, we can generally only recommend the permutation Wald-type tests for both designs and the quantile regression approach for the 2 × 2 layout. All other methods show a quite (partially extremely) liberal or conservative behavior for a significant number of settings. In the next subsection, we analyze how the latter affects the power performances under shift alternatives.
**Table 2.** Type-I error rates in % for the $2 \times 3$ design.

| Censoring | Low | Medium I | Medium II |
|-----------|-----|----------|-----------|
|           |  Two-sided | One-sided | Two-sided | One-sided | Two-sided | One-sided |
| Distr     |       |          |           |          |           |          |
| Exp       |       |          |           |          |           |          |
| $n_1$     | 5.0  | 3.5      | 5.3       | 8.8      | 10.6      | 5.1       | 4.6       | 5.1       | 7.7       | 11.7      | 5.6       | 2.6       | 5.1       | 10.6      | 10.1      |
| $2n_1$    | 4.8  | 3.3      | 4.5       | 6.9      | 9.7       | 5.7       | 3.7       | 5.4       | 6.6      | 10.4      | 5.2       | 3.6       | 4.8       | 8.1       | 9.7       |
| $n_4$     | 4.2  | 3.4      | 4.8       | 8.0      | 10.7      | 5.2       | 4.4       | 5.0       | 7.5      | 12.8      | 5.5       | 2.8       | 4.6       | 10.3      | 11.1      |
| $2n_4$    | 5.1  | 3.0      | 4.9       | 7.1      | 11.4      | 5.8       | 4.1       | 4.9       | 6.6      | 12.2      | 5.0       | 3.2       | 5.3       | 8.5       | 11.7      |
| Weib      |       |          |           |          |           |          |
| $n_1$     | 5.4  | 3.2      | 5.4       | 4.3      | 10.4      | 5.4       | 3.3       | 5.3       | 3.6      | 11.6      | 5.1       | 2.9       | 5.0       | 5.5       | 11.3      |
| $2n_1$    | 5.1  | 3.6      | 5.1       | 4.1      | 10.0      | 5.2       | 3.4       | 5.2       | 4.0      | 10.0      | 4.9       | 3.7       | 5.2       | 4.7       | 8.7       |
| $n_4$     | 4.7  | 2.9      | 5.6       | 4.5      | 11.6      | 5.3       | 3.6       | 5.4       | 3.8      | 11.1      | 5.2       | 3.2       | 5.2       | 5.9       | 10.1      |
| $2n_4$    | 5.0  | 3.4      | 5.3       | 4.1      | 10.4      | 5.1       | 3.5       | 5.3       | 4.0      | 11.4      | 5.2       | 4.3       | 5.4       | 5.2       | 11.5      |
| LogN      |       |          |           |          |           |          |
| $n_1$     | 5.1  | 3.2      | 4.8       | 8.4      | 10.0      | 4.8       | 3.1       | 5.1       | 7.5      | 10.7      | 5.3       | 1.9       | 5.1       | 10.8      | 10.2      |
| $2n_1$    | 5.2  | 2.8      | 4.7       | 7.1      | 9.3       | 5.6       | 3.5       | 5.7       | 7.5      | 10.7      | 5.1       | 2.8       | 5.1       | 8.7       | 9.5       |
| $n_4$     | 4.5  | 2.6      | 4.7       | 8.3      | 10.3      | 5.5       | 4.3       | 5.0       | 7.6      | 11.2      | 5.3       | 1.9       | 5.3       | 10.9      | 11.1      |
| $2n_4$    | 4.9  | 2.8      | 5.2       | 7.8      | 10.6      | 6.0       | 3.6       | 4.9       | 6.8      | 11.3      | 5.4       | 2.9       | 5.3       | 9.1       | 11.4      |
| Mix       |       |          |           |          |           |          |
| $n_1$     | 6.4  | 3.4      | 6.0       | 7.8      | 11.0      | 6.1       | 4.1       | 5.1       | 5.8      | 12.1      | 6.6       | 2.6       | 5.7       | 9.8       | 11.6      |
| $2n_1$    | 6.0  | 3.5      | 6.0       | 6.5      | 10.4      | 6.4       | 3.8       | 6.1       | 6.1      | 11.0      | 5.1       | 3.2       | 5.7       | 7.3       | 9.6       |
| $n_4$     | 6.0  | 3.3      | 5.6       | 6.9      | 11.7      | 6.0       | 3.6       | 5.3       | 5.8      | 13.2      | 5.7       | 2.3       | 4.8       | 8.4       | 11.9      |
| $2n_4$    | 5.2  | 2.9      | 5.7       | 6.4      | 12.3      | 6.9       | 4.3       | 5.9       | 6.1      | 12.7      | 5.2       | 3.1       | 5.2       | 7.6       | 11.7      |
| Largest   | .35  | .26      | .34       | .40      | .46       | .36       | .30       | .33       | .38      | .48       | .35       | .29       | .33       | .44       | .46       |

Notes: Type-I error rate in % (nominal level $a = 5\%$) for testing the null hypothesis of no main effect of the first factor A in the $2 \times 3$ design for the asymptotic (Asy) and permutation Wald-type tests based on the one-sided and two-sided interval variance estimators as well as for the quantile regression method (QR).

Values inside the 95% binomial interval $[4.4, 5.6]$ are printed bold. The largest standard error (s.e.) per column was derived by taking the largest type-I error rate in the column as the true type-I error rate.

### 5.4 Power behavior under shift alternatives

This subsection is devoted to the comparison of the tests’ power performances under shift alternatives, which were introduced at the beginning of this section. We ran the simulations for the larger sample size settings, i.e. $n = 2n_j$ for $j = 1, \ldots, 4$. Since the results for the balanced and unbalanced scenarios lead to the same overall conclusions, we only display the results for the respective unbalanced scenarios. In Figures 1 and 2, the power curves for the $2 \times 2$ design testing for no main effect and no interaction effect, respectively, are presented for all five considered tests. Except for testing for no main effect under the Weibull distribution, the power values of the asymptotic Wald-type test with the one-sided approach show the highest values.

In particular, for the low censoring settings, the differences to its competitors are most pronounced. This can be explained by the partially very extremely liberal behavior under the null hypotheses, which we observed in the previous subsection. Consequently, the comparison with this specific asymptotic test is not fair at all and needs to be taken with a pinch of salt. On the other hand, for the two-sided interval variance strategy, it can be seen that the conservative behavior of the asymptotic tests also affects the power performance: compared to the more accurate permutation approach, they show a significant loss of power. When comparing both permutation tests, it can be seen that the one-sided strategy leads to higher power values in all settings except under Weibull distributions, where the two respective power curves are nearly indistinguishable.

We finally turn to the most important comparison with the jackknife approach for the quantile regression. A study of Figure 1 for the main effect set-up reveals that the curves for the jackknife method are slightly above the ones for the permutation tests under all Weibull settings and for the combination of low censoring and exponentially distributed survival times. However, it should be emphasized that for these four scenarios, the observed type-I error rates of the jackknife method were slightly liberal with values around 6.5%, while the permutation tests reliably kept the nominal level. For the remaining cases, the curves of the jackknife method and the permutation tests with the one-sided interval strategy are almost identical.

These conclusions change completely when testing for interaction effects (Figure 2). Here, the permutation test with the one-sided interval strategy shows the overall best power behavior while the jackknife quantile regression approach exhibits the lowest power in most situations.
Finally, the impressions for the $2 \times 3$ design (Figure 3) are similar to the $2 \times 2$ design for all Wald-type tests. However, a significant difference can be observed for the jackknife method: it now shows much flatter power curves. The latter is rather surprising in view of its liberal type-1 error performance under the null hypothesis.

5.5 Recommendations

To summarize the simulation results, we recommend the permutation test with the one-sided interval strategy over the other three Wald-type approaches, as it controls the type-1 error most accurately. It also leads to the highest power values among all Wald-type tests controlling the nominal level $z$. A direct comparison of this permutation Wald-type test with the jackknife quantile regression approach clearly favors the Wald-type test when testing for no interaction effect in the $2 \times 2$ design and for no main effect in $2 \times 3$ layouts. When testing for no main effect in the $2 \times 2$ setting, both approaches can be recommended.

6 Illustrative data analysis

We illustrate the use of the developed tests by re-analyzing a controlled clinical trial conducted in Denmark during the period 1962–1974. Included patients suffered from a histologically verified liver cirrhosis and the times until death have been recorded. Due to right-censoring, these times were not always observable. The study aimed to analyze the effectiveness of a treatment with prednisone against placebo with respect to survival, while also the influence of several prognostic variables is assessed. The impact of these additional variables on the survival chances was analyzed by means of a Cox proportional hazards model. In addition, the authors pointed out on page 892 that “Another measure for the prognosis calculated from PI (prognostic index, comment by the authors) is the MST indicating the span of time that the patient will survive with 50% probability.” They also provided a
plot of the MST against PI. We complement this descriptive analysis of the MST with inferential investigations based upon our newly developed nonparametric methods.

In the following, the dataset csl available through the R-package \textit{timereg} is considered; it contains a subset of size 446 of the original data. After an initial analysis of the factors, treatment and sex, we will investigate the influence of the two variables, treatment and prothrombin level (at baseline), on the survival median for two specific subsets of the data. A prothrombin level of <70\% of the normal level is considered abnormal (Andersen et al., p. 33).\footnote{Andersen et al.}

The data sets are analyzed with the help of the five tests that were compared in the simulation study in the previous section, i.e. the four Wald-type tests developed in this paper (asymptotic/permutation; one-/two-sided interval variance estimator) and the quantile regression method applied to the median. Ties in the event times have been broken with the help of small random noise. For all tests, we choose the significance level $\alpha = 5\%$.

6.1 Analysis of the complete dataset

We begin with the analysis of the complete dataset to find out whether there are main and/or interaction effects between the factors treatment and sex on the MSTs. Based on these outcomes, we will later divide the dataset for a more detailed analysis. Table 3 summarizes a few characteristics of the complete dataset. The sample sizes are quite imbalanced regarding sex; more men than women were included. Also the censoring rates are higher for the female subgroups which could be caused by a lower mortality rate. The age and prothrombin levels of the individuals at baseline are quite similar which is in line with the fact that patients have been randomized into both treatment groups.

The estimated MSTs (Table 3) and the plots in Figure 4 indicate a possible interaction effect between sex and treatment on the MST. While the medians in both male treatment groups are very close (4.43 and 4.37 years), the

![Figure 2. Power curves for the $2 \times 2$ design testing for no interaction effect of the asymptotic (Asy) and permutation (Per) Wald-type tests using the two-sided and one-sided interval variance estimators and of the quantile regression method for $n = 2n_2$ and shift alternatives $d = (0, 0.0)$. The largest standard error per simulated power is 0.7\%.](image)
medians of the female groups differ quite a lot (3.20 and 6.74 years). A statistical investigation of the MST shall clarify the question of whether there is a significant difference. The results of all hypothesis tests can be found in Table 4. While there is no significant gender effect, only the Wald-type tests based on the one-sided variance estimators found a significant treatment effect. The permutation version of this test was the preferred one according to our simulation study above. For the tests based on the two-sided variance estimator, we obtain borderline significant results ($p$-values: .06 and .051). An interaction effect between treatment and sex has been revealed by all applied Wald-type tests but not by the quantile regression ($p$-value: .095). Thus, the results of the Wald-type tests are in line with what we expected from our initial numerical and graphical analyses. By the way, the original paper\textsuperscript{40} reported a significant sex effect but a non-significant treatment effect within the Cox model.

Figure 3. Power curves for the $2 \times 3$ design testing for no main effect of the asymptotic (Asy) and permutation (Per) Wald-type tests using the two-sided and one-sided interval variance estimators and of the quantile regression method for $n = 2n_4$ and shift alternatives $\delta = (\delta, \delta, 0, 0, 0, 0)$. The largest standard error per simulated power is 0.7%.

Table 3. Summary of the complete dataset according to sex and treatment.

| Sex  | Treatment | Male | Female | Overall |
|------|-----------|------|--------|---------|
|      |           | Placebo | Prednisone | Placebo | Prednisone | Placebo | Prednisone | Overall |
| Sample size | 125 | 132 | 95 | 94 | 446 |
| Censoring rate (%) | 36.0 | 35.6 | 37.9 | 51.1 | 39.5 |
| Av. age (years) | 57.9 (10.0) | 58.6 (8.8) | 67.1 (9.1) | 65.9 (11.3) | 62.9 (11.3) | 58.9 (11.3) | 59.2 |
| Av. prothrombin level (% of normal) | 70.8 (21.1) | 68.5 (22.1) | 70.2 (25.8) | 67.1 (23.8) | 70.2 (25.8) | 69.2 |
| Est. median survival time (years) | 4.43 (0.61) | 4.37 (0.74) | 3.20 (1.07) | 6.74 (1.05) | 4.78 |

Notes: The numbers in parentheses give the estimated standard deviations: for age and prothrombin level it is the sample standard deviation and the standard deviation of the estimated median survival time is estimated by $\hat{\sigma}^\text{two-sided}$. 

The medians of the female groups differ quite a lot (3.20 and 6.74 years). A statistical investigation of the MST shall clarify the question of whether there is a significant difference. The results of all hypothesis tests can be found in Table 4. While there is no significant gender effect, only the Wald-type tests based on the one-sided variance estimators found a significant treatment effect. The permutation version of this test was the preferred one according to our simulation study above. For the tests based on the two-sided variance estimator, we obtain borderline significant results ($p$-values: .06 and .051). An interaction effect between treatment and sex has been revealed by all applied Wald-type tests but not by the quantile regression ($p$-value: .095). Thus, the results of the Wald-type tests are in line with what we expected from our initial numerical and graphical analyses. By the way, the original paper\textsuperscript{40} reported a significant sex effect but a non-significant treatment effect within the Cox model.
However, one should be aware that the Cox model measures different effects. Moreover, the original study did not consider possible interaction effects and also included additional covariates.

### 6.2 Subset analysis of the female group

Because of the significant influence of sex in the original Cox analysis\(^{40}\) and the significant treatment–sex interaction effect in the complete data analysis, we decided to split the dataset according to gender. At this stage of the analysis, we only focus on the subset of females. For these, we are going to test for main and interaction effects in the factors treatment and normal/abnormal prothrombin level. A first glance at the summaries in Table 5 reveals that there are quite big differences in the censoring rates and estimated survival medians depending on normal or abnormal prothrombin levels. It seems that female patients with normal prothrombin values have a higher chance...
of survival; see Figure 5. This would be in line with the significant protective effect of normal prothrombin levels that were found for the complete dataset in the original Cox analysis.40

The results of all hypothesis tests can be found in Table 6. All Wald-type tests agree with a significant pro-thrombin effect. The quantile regression, on the other hand, shows a small but still non-significant $p$-value (.083).

In addition, most Wald-type tests and also the quantile regression could not detect a treatment effect. This is also in line with the result of the overall Cox analysis.40 In addition, none of the tests applied here revealed a treatment–prothrombin interaction effect on the MSTs.

6.3 Subset analysis of males aged 60–69 years

In the original Cox analysis,40 the variable age was highly significant. That is why we decided to use a rather homogenous age group for this subset analysis. Moreover, since the proposed permutation test can reliably handle small sample sizes, we preferred a rather compact time range for this group. This lead to our final choice of a 10-year time span and, specifically, to the subgroup of male patients who were between 60 and 69 years old at baseline. The related Kaplan–Meier curves are displayed in Figure 6. The plots and the summary in Table 7 point out a possible prothrombin effect. The sample sizes in all four groups of males divided according to treatment and (ab)normal prothrombin levels are quite small and imbalanced (14, 21, 27, 32); note that this sample size and censoring setting corresponds to our simulation scenarios with $n = 2n_3$ and $cr = cr_1$ (small differences are due to rounding). Yet, all applied tests (Table 8) are powerful enough to confirm the presence of a prothrombin effect. On the other hand, no test found a significant treatment or interaction effect.

7 Summary and discussion

Factorial designs with time-to-event endpoints are usually analyzed by Cox regression models. However, quantification of treatment or interaction effects by a single parameter can only be warranted under proportional hazards. This gives rise to study other effect sizes than hazard ratios in factorial designs. In this paper, we considered MSTs as they are frequently reported, easy to understand,13 and have therefore been propagated
recently. Furthermore, in addition to a mere reporting of survival chances at specific time points of clinical interest, medians add valuable information: survival medians are quantities given in units of time and they provide an estimate that is of personal interest to many patients with a severe disease. Also, other survival quantiles may be used to answer the often essential question “How much time do I have left?”

Our main aim was to demonstrate that the analysis of MSTs is feasible not just within two- or k-sample settings but even within general factorial designs, in particular, allowing for inference on interaction effects. Moreover, contrary to the established quantile regression, no multiple dummy variables for nominal factors (e.g. different treatments) need to be introduced within the factorial design setting. This property is even favorable in uncensored situations. To this end, we proposed a flexible new class of permutation procedures that can be applied to test main and interaction effects or more general contrasts formulated in terms of MSTs. They are based on Wald-type statistics that ensure a robust behavior against non-exchangeable and more complex heterogeneous settings. In fact, we proved consistency of the new methodology under weak conditions; thereby also revealing new insights on permutation empirical processes and variance estimations based upon one-sided confidence intervals which can be found in the technical supplementary material.

---

Figure 6. Male aged 60–69 years subset, divided according to treatment: prednisone (left panel) and placebo: (right panel). Each panel displays two Kaplan–Meier curves that correspond to normal/abnormal levels of prothrombin.

Table 7. Summary of the male aged 60–69 years subset according to treatment and prothrombin level at baseline.

| Treatment      | Placebo | Prednisone |
|----------------|---------|------------|
| Prothrombin level | < 70    | ≥ 70       |
| Sample size     | 14      | 27         |
| Censoring rate (%) | 7.1    | 29.6       |
| Av. age (years) | 65.4 (2.6) | 64.4 (2.8) |
| Av. prothrombin level (% of normal) | 49.7 (14.1) | 87.4 (11.9) |
| Est. median survival time (years) | 2.06 (1.34) | 4.42 (0.87) |

Overall sample size: 94

Notes: The numbers in parentheses give the estimated standard deviations: for age and prothrombin level it is the sample standard deviation and the standard deviation of the estimated median survival time is estimated by $\hat{\sigma}_{\text{two-sided}}$.

Table 8. p-Values of all tests applied to the male aged 60–69 years subset with factors treatment and prothrombin.

| Test procedure | Wald-type test | QR |
|----------------|---------------|----|
| Variance estimation | Two-sided | One-sided | Jackknife |
| Approximation method | Asymptotic | Permutation | Asymptotic | Permutation | Jackknife |
| Main effect treatment | .673 | .657 | .624 | .624 | .972 |
| Main effect prothrombin | .019 | .014 | .014 | .007 | .025 |
| Interaction effect | .756 | .743 | .714 | .717 | .846 |

Note: Significant p-values (at level $\alpha = 5\%$) are printed in bold type.
Beyond theoretical investigations, illustrative data analyses and exhaustive simulations for different designs and various distributional settings supported the usage of the new approach. In particular, a permutation test together with a so-called one-sided variance estimation implementation yielded more reliable results than existing methods in our simulation study. Moreover, the presented analysis of a real dataset on liver cirrhosis patients confirmed these findings: the proposed permutation Wald-type test is quite powerful and can lead to rejections of null hypotheses even for small sample sizes in combination with moderate censoring rates.

In addition to inference, our tests’ outcomes could also be used for factor selection, so that more precise inference or confidence regions for (contrasts of) the MSTs of more influential factors can be obtained.

In the future, we plan to transfer the recent findings to ratios of MST contrasts. Moreover, similar extensions of other time-to-event effect sizes, such as the (restricted) mean survival, to factorial designs will be investigated. In this context, extensions of the permutation methodology to meta-analysis studies based on medians or the restricted mean survival time will also be part of future research.

To guarantee a simple application of the presented methods, we currently work on their implementation into the R-package being available on CRAN soon and a corresponding Shiny web application. The respective R-function is coined , an abbreviation for median survival analysis of variance.

Acknowledgements
The authors are grateful to the editor, the associate editor, and the two referees for their comments that improved the paper’s quality. The authors thank Stefan Inerle for computational assistance.

Data availability statement
Data sharing is not applicable to this article as no new data were created or analyzed in this study.

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: Marc Ditzhaus and Markus Pauly were funded by the Deutsche Forschungsgemeinschaft (grant no. PA-2409 5-1).

Supporting information
All proofs and technical derivations are available as supporting information.

ORCID iDs
Marc Ditzhaus https://orcid.org/0000-0001-9235-1905
Dennis Dobler https://orcid.org/0000-0002-9040-0854

References
1. Arboretti R, Fontana R, Pesarin F, et al. Nonparametric combination tests for comparing two survival curves with informative and non-informative censoring. Stat Methods Med Res 2018; 12: 3739–3769.
2. Brendel M, Janssen A, Mayer CD, et al. Weighted logrank permutation tests for randomly right censored life science data. Scand J Stat 2015; 41: 742–761.
3. Ditzhaus M and Pauly M. Wild bootstrap logrank tests with broader power functions for testing superiority. Comp Statist Data Anal 2019; 136: 1–11.
4. Ditzhaus M and Friedrich S. More powerful logrank permutation tests for two-sample survival data. J Stat Comput Simul 2020; 90: 2209–2227.
5. Ditzhaus M, Janssen A and Pauly M. Permutation inference in factorial survival designs with the CASANOVA. arXiv:2004.10818 2020.
6. Gorfine M, Schlesinger M and Hsu L. k-sample omnibus non-proportional hazards tests based on right-censored data. Stat Methods Med Res 2020; 29: 2830–2850.
7. Ellis LM, Bernstein DS, Voest EE, et al. American society of clinical oncology perspective: raising the bar for clinical trials by defining clinically meaningful outcomes. J Clin Oncol 2014; 32: 1277–1280.
8. Brückner M and Brannath W. Sequential tests for non-proportional hazards data. *Lifetime Data Anal* 2017; 23: 339–352.
9. Kalbfleisch JD and Prentice RL. Estimation of the average hazard ratio. *Biometrika* 1981; 68: 105–112.
10. Dobler D and Pauly M. Bootstrap-and permutation-based inference for the Mann–Whitney effect for right-censored and tied data. *TEST* 2018; 27: 639–658.
11. Dobler D and Pauly M. Factorial analyses of treatment effects under independent right-censoring. *Stat Methods Med Res* 2020; 29: 325–343.
12. Koziol JA and Jia Z. The concordance index C and the Mann–Whitney parameter Pr(X > Y) with randomly censored data. *Biom J* 2009; 51: 467–474.
13. Ben-Aharon O, Magnezi R, Leishno M, et al. Median survival or mean survival: which measure is the most appropriate for patients, physicians, and policymakers? *Oncologist* 2019; 24: 1469.
14. Royston P and Parmar MKB. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Med Res Methodol* 2013; 13: 152.
15. Brookmeyer R and Crowley J. A confidence interval for the median survival time. *Biometrics* 1982; 38: 29–41.
16. Brookmeyer R and Crowley J. A k-sample median test for censored data. *J Amer Statist Assoc* 1982; 77: 433–440.
17. Chen Z and Zhang G. Comparing survival curves based on medians. *BMC Med Res Methodol* 2016; 16: 33.
18. Martinussen T and Pipper CB. Estimation of odds of concordance based on the Aalen additive model. *Lifetime Data Anal* 2013; 19: 100–116.
19. Green S. Factorial designs with time to event endpoints. In: Crowley J and Ankerst DP (eds) *Handbook of statistics in clinical oncology*. Boca Raton, FL: Chapman and Hall/CRC, 2005, pp.201–210.
20. Chen YI and Chang YM. Covariates-dependent confidence intervals for the difference or ratio of two median survival times. *Stat Med* 2007; 26: 2203–2213.
21. Su JQ and Wei LJ. Nonparametric estimation for the difference or ratio of median failure times. *Biometrics* 1983; 49: 603–607.
22. Ditzhaus M, Fried R and Pauly M. QANOVA: quantile-based permutation methods for general factorial designs. *arXiv:1912.09146v2* 2019.
23. Chung EY and Romano JP. Exact and asymptotically robust permutation tests. *Ann Statist* 2013; 41: 484–507.
24. Janssen A and Pauls T. How do bootstrap and permutation tests work? *Ann Statist* 2013; 3: 768–806.
25. Pauly M, Brunner E and Konketscheke F. Asymptotic permutation tests in general factorial designs. *J R Stat Soc Ser B Stat Methodol* 2015; 77: 461–473.
26. Brunner E, Dette H and Munk A. Box-type approximations in nonparametric factorial designs. *J Amer Statist Assoc* 1997; 92: 1494–1502.
27. Smaga L. Diagonal and unscaled wald-type tests in general factorial designs. *Electron J Stat* 2017; 11: 2613–2646.
28. Sander JM. The weak convergence of quantiles of the product-limit estimator. Stanford University, Department of Statistics, Technical report number 5, 1975.
29. McKeans J and Schrader RG. A comparison of methods for studentizing the sample mean. *Commun Statist B* 1984; 13: 751–773.
30. Price RM and Bonett DG. Estimating the variance of the sample median. *J Stat Comput Simul* 2001; 68: 295–305.
31. Chakraborti S. Large sample tests for equality of medians under unequal right-censoring. *Comm Statist Theory Methods* 1988; 17: 4075–4084.
32. Chakraborti S. A class of tests for homogeneity of quantiles under unequal right-censorship. *Stat Probab Lett* 1990; 9: 107–109.
33. Aalen O. Nonparametric inference in connection with multiple decrement models. *Scand J Stat* 1976; 3: 15–27.
34. Hemnerik J and Goeman J. Exact testing with random permutations. *TEST* 2018; 27: 811–825.
35. Pauly M, Asendorf T and Konketscheke F. Permutation-based inference for the AUC: a unified approach for continuous and discontinuous data. *Biometrical J* 2016; 58: 1319–1337.
36. Portnoy S. Censored quantile regression. *J Amer Statist Assoc* 2003; 98: 1001–1012.
37. Portnoy S. The jackknife’s edge: inference for censored regression quantiles. *Comput Statist Data Anal* 2014; 72: 273–281.
38. Koenker R, Portnoy S, Ng PT, et al. A package for quantile regression in R, Version 5.55. https://CRAN.R-project.org/package=quantreg (2020, accessed 7 December 2020).
39. R Core Team. R: a language and environment for statistical computing, https://R-project.org/ (2020, accessed 7 December 2020).
40. Schlichting P, Christensen E, Andersen PK, et al. Prognostic factors in cirrhosis identified by Cox’s regression model. *Hepatology* 1983; 3: 889–895.
41. Andersen PK, Borgan Ø, Gill RD, et al. *Statistical models based on counting processes*. New York, NY: Springer, 1993.
42. Green S, Liu P-Y and O’Sullivan J. Factorial design considerations. *J Clin Oncol* 2002; 20: 3424–3430.
43. Michiels S, Piedbois P, Burdett S, et al. Meta-analysis when only the median survival times are known: a comparison with individual patient data results. *Int J Technol Assess* 2005; 21: 119–125.
44. Wei Y, Royston P, Tierney JF, et al. Meta-analysis of time-to-event outcomes from randomized trials using restricted mean survival time: application to individual participant data. *Stat Med* 2015; 34: 2881–2898.