Factorial analyses of treatment effects under independent right-censoring

Dennis Dobler and Markus Pauly

Abstract
This paper introduces new effect parameters for factorial survival designs with possibly right-censored time-to-event data. In the special case of a two-sample design, it coincides with the concordance or Wilcoxon parameter in survival analysis. More generally, the new parameters describe treatment or interaction effects and we develop estimates and tests to infer their presence. We rigorously study their asymptotic properties and additionally suggest wild bootstrapping for a consistent and distribution-free application of the inference procedures. The small sample performance is discussed based on simulation results. The practical usefulness of the developed methodology is exemplified on a data example about patients with colon cancer by conducting one- and two-factorial analyses.

Keywords
Factorial designs, Kaplan–Meier estimator, nonparametric statistics, quadratic forms, wild bootstrap

1 Motivation and introduction
Factorial designs are often encountered in biomedical and clinical trials. Here, not only the (main) effects of separate factors but also interaction effects that are related to possibly complex factor combinations are of importance. Such interaction effects may even alter the interpretation of main effects leading to the established comment that ‘it is desirable for reports of factorial trials to include estimates of the interaction between the treatments’.

On the other hand, nonparametric estimation and the inference of adequate effects in such designs can be rather involved. In particular, most existing inference procedures have focused on testing hypotheses formulated in terms of distribution functions. But those cannot be inverted to obtain confidence intervals or regions for meaningful effects. Only recently, nonparametric methods for inferring adequate effects in general factorial designs with independent and dependent observations have been established. These procedures are, however, only developed for completely observed data and not applicable for partially observed time-to-event data. Since many clinical studies are concerned with survival outcomes, adequate statistical inference methods for complex factorial time-to-event designs are of particular interest.

To detect main effects, weighted logrank tests or their extensions may be applied in case of two or multiple samples. However, these procedures only infer conclusions in terms of cumulative hazard functions and cannot be applied to obtain concrete effect parameters with informative confidence intervals nor tests for the presence of interactions. In practice, interaction effects are usually modeled with the help of Cox-, Aalen-, or even Cox-Aalen regression models with factors as covariates and incorporated interaction terms. However, although very flexible, these models are usually more driven towards hazards modeling from continuous event times. Moreover, the incorporation of several factor variables (e.g., via multiple dummy variables per factor) can become cumbersome; this is so even in the uncensored case, especially when interactions are incorporated.

The above problems directly motivate a nonparametric approach for estimating and inferring main and interaction effects in factorial designs with censored observations. So far, only few nonparametric survival procedures exist in this context. They are based on a purely nonparametric model that does not require

1Department of Mathematics, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands
2Institute of Statistics, Ulm University, Ulm, Germany

Corresponding author:
Dennis Dobler, Faculty of Sciences, Department of Mathematics, Vrije Universiteit Amsterdam, Amsterdam 1081 HV, the Netherlands.
Email: d.dobler@vu.nl
any multiplicative or additive structure of the hazards and can even be applied for arbitrary, possibly noncontinuous survival distributions (i.e. it can be readily used for survival times rounded to days, weeks or months). Moreover, it leads to tests for main and interaction effects in case of independent right-censored data. However, these tests suffer from several drawbacks: the procedure is based on a rather strong assumption on the underlying censorship distribution which is often hard to verify in practical situations. In addition, null hypotheses are only formulated in terms of distribution functions. As a result, there is no direct quantification and estimation of main and interaction effects in terms of confidence intervals as, e.g. required by regulatory authorities (ICH E9 Guideline, 1998, p. 25)\textsuperscript{35}

This is to be changed in the current paper. We develop and rigorously analyze nonparametric inference procedures, i.e. tests and confidence intervals, for meaningful effect sizes in factorial survival designs, where data may be subject to random right-censoring.

Similar to the adaption of the test by Brunner and Munzel test to the two-sample survival set-up by Dobler and Pauly, we consider the recently proposed unweighted nonparametric effects of Brunner et al. and extend their ansatz to a general survival setting.\textsuperscript{17,36,37} In the special case of proportional hazards, these effects have a direct relationship to hazard ratios in two-sample settings.\textsuperscript{38} At the same time, they remain meaningful in case of non-proportional hazards. This fact makes the effect sizes even more appealing for practical purposes.

In the following sections, the statistical model and important results on the basic estimators are presented and the resulting test statistic for the null hypotheses of interest is stated and mathematically analyzed. Since the asymptotic distribution of the test statistic depends on unknown parameters, we propose a distribution-free multiplier resampling approach and prove its consistency. Next, it is supplemented by a simulation study to assess the finite sample properties of the proposed procedure. It is further exemplified on a colon cancer study, where in the original study the analysis was made in terms of Cox models.\textsuperscript{39} Finally, the paper closes with concluding comments. All proofs are deferred to the technical Appendix 1.

2 The set-up

To establish the general model, we consider sequences of mutually independent random variables

\[ T_{ik} \overset{\text{ind}}{\sim} S_i \quad \text{and} \quad C_{ik} \overset{\text{ind}}{\sim} G_i \quad (i = 1, \ldots, d, \ k = 1, \ldots, n_i) \]  

(1)

where \( T_{ik} \) denotes the actual survival time of subject \( k \) in group \( i \) and \( C_{ik} \) the corresponding censoring variable. Moreover, to even allow for ties or survival times rounded to weeks or months, the survival functions \( S_i \) and \( G_i, i = 1, \ldots, d, \) defined on \((0, \infty)\) may be possibly discontinuous. That is, the corresponding hazard rates may, but need not exist. The actually observable data consist of the right-censored survival times \( X_{ik} = T_{ik} \wedge C_{ik} \) and the uncensoring indicators \( \delta_{ik} = 1[T_{ik} \leq C_{ik}], \ i = 1, \ldots, d, \ k = 1, \ldots, n_i. \) In this set-up, a factorial structure can be incorporated by splitting up indices, see, e.g., section 6 for an exemplary \( 2 \times 3 \) two way layout.

In the special case of \( d = 2 \) groups with continuous survival times, an estimator of the concordance probability

\[ w = P(T_{11} > T_{21}) = -\int S_1 dS_2 \]

that a randomly chosen subject from the first group survives longer than someone from the second group who had been introduced.\textsuperscript{40} If all subjects are completely observable, this effect size \( w \) is directly estimable through empirical distribution functions. In this case, it reduces to the well-known Mann–Whitney effect underlying the Brunner and Munzel test.\textsuperscript{36} In the present right-censored case, however, the right tails of the survival times’ distributions, and thus \( w \), are usually not identifiable. We are going to resolve this issue by introducing an identifiable quantity \( \tilde{w} \), which is related to \( w \), in equation (2) below. Anyhow, inference procedures for \( w \) and related quantities in survival set-ups (such as the concordance parameter or the average hazard ratio) have been developed as well.\textsuperscript{17,38} However, an extension of the definition of \( w \) to the more general design (1), allowing for an arbitrary factorial structure, is not straightforward. In particular, for the case of completely observed data, several pitfalls which possibly lead to paradoxical results when working with a ‘wrong’ extension of \( w \) have been pointed out.\textsuperscript{17,41}

Adopting the solution of Brunner et al. to the present situation, we introduce an additional ‘benchmark’
survival time $Z$, independent of the above, with averaged survival function $Z \sim \tilde{S} = \frac{1}{d} \sum_{i=1}^{d} S_i$. This is used to extend $w$ to

$$\tilde{p}_i = P(T_i > Z) + \frac{1}{2} P(T_i = Z) = -\int S_i^+ d\tilde{S}$$

where the superscript $\pm$ denotes the average of a right-continuous function and its left-continuous version. The use of such normalized survival functions adequately handles discrete components of the survival distribution, i.e. ties in the data are explicitly allowed.

The choice of the effect parameter $\tilde{p}_i$ is motivated by recent findings on nonparametric analyses of factorial designs with complete observations.\textsuperscript{17,41} Therein, it is stressed that other choices, e.g., pair-wise comparisons of all concordance probabilities $w$ or comparisons with the weighted survival function $\sum_{j=1}^{d} S_j$ instead of $\tilde{S}$, may easily result in paradoxical outcomes. This is no issue for the effects $\tilde{p}_i$ which are sample size independent. For later calculations, we emphasize that the effect parameters are balanced in the mean. In particular, we have

$$\frac{1}{d} \sum_{i=1}^{d} \tilde{p}_i = -\int S_i^+ d\tilde{S} = \frac{1}{2} \quad \text{and} \quad \tilde{p}_i = -\sum_{j \neq i} S_j^+ dS_j + \frac{1}{2d}$$

From a practical point of view, estimation of the $\tilde{p}_i$'s would need ‘arbitrarily’ large survival times since the integral is defined on $(0, \infty)$. However, every study ends at a certain point in time. For practical applicability, we therefore assume that the censoring times are bounded and we have to modify the $\tilde{p}_i$'s accordingly: denote by $\tau > 0$ the smallest out of the largest possible censoring time per group, or any smaller value. In comparisons of survival times, which belong to different groups and which exceed $\tau$, no group shall be favored. In other words, the remaining mass has to be split up equally among the groups. Technically, this is realized by setting the remaining mass of the survival functions to zero: $S_i(\tau) = 0$. Redefining $S_i$ and $\tilde{S}$ from now on as the survival functions of $\min(T_i, \tau)$ and $\min(Z, \tau)$, respectively, this translates into the nonparametric concordance effects

$$p_i = P(\min(T_i, \tau) > \min(Z, \tau)) + \frac{1}{2} P(\min(T_i, \tau) = \min(Z, \tau)) = -\int S_i^+ d\tilde{S} \quad (2)$$

Obviously, all of the above-discussed positive properties of the effects parameter $\tilde{p}_i$ also transfer to the nonparametric concordance effects $p_i$: it is a meaningful effect measure for ordinal and metric data, sample size independent, and allows for a suitable treatment of ties.

We aggregate all effects into the vector $p = (p_1, \ldots, p_d)'$ and borrow a trick\textsuperscript{17,42} to express them as

$$p = \left( I_d \otimes \frac{1}{d} I_d \right) \cdot (w_1, \ldots, w_d)' =: E_d \cdot w$$

Here, $w_i = (w_{i1}, \ldots, w_{id})'$ is the $d$-vector of effects for direct comparisons of group $i$ with respect to all groups $j = 1, \ldots, d$, and $S = (S_1, \ldots, S_d)'$ is the aggregation of all survival functions. Moreover, $I_d$ denotes the identity matrix in $\mathbb{R}^d$ the $d$-dimensional vector of 1's and the symbol $\otimes$ denotes the Kronecker product. In this way, the $i$th entry of $w_i$ is $w_{ii} = \frac{1}{d}$ which makes sense because equal groups should be valued equally high. Anyhow, equation (3) shows that the problem of estimating $p$ reduces to the estimation of the pair-wise effects $w_{ij}$. But this can be achieved by substituting each involved survival function $S_i$ by its Kaplan–Meier estimator $\hat{S}_i$, $i = 1, \ldots, d$.\textsuperscript{43}

Proceeding in this way, we denote by $\hat{w}$ and $\hat{w}_i$ these estimated counterparts of $w$ and $w_i$. Let $N = \sum_{i=1}^{d} n_i$ be the total sample size. Below we establish the asymptotic normality of $\sqrt{N}(\hat{w} - w)$ under the following framework

$$N^{-1} n := \left( \frac{n_1}{N}, \ldots, \frac{n_d}{N} \right)' \to \lambda := (\lambda_1, \ldots, \lambda_d)' \in (0, 1)^d \quad (4)$$

as $\min n \to \infty$. To give a detailed description of the resulting asymptotic covariance structure, however, we first have to introduce some additional notation: Let $D[0, \tau]$ be the space of all càdlàg-functions on $[0, \tau]$, equipped with the Skorokhod metric, and $BV[0, \tau] \subset D[0, \tau]$ its subspace of càdlàg-functions with bounded variation.
For the subsequent arguments, it is essential that we can represent \( w = \phi \circ S \) as a functional of \( S \). In particular, the functional

\[
\phi : (BV[0, r])^d \to \mathbb{R}^d, \quad (f_1, \ldots, f_d) \mapsto (-\int f_{ij}^d)_{i,j=1}^d
\]

with inner index \( j \), is Hadamard-differentiable at \( S \); see the proof of Lemma 1 below for details. We denote its Hadamard-derivative at \( S \) by \( d\phi_w \), which is a continuous linear functional. For technical reasons, we assume throughout that \( P(T_{i1} > r) > 0 \) for all groups \( i = 1, \ldots, d \). We may now state the first preliminary but essential convergence result.

**Lemma 1.** Under the asymptotic regime (4) we have

\[
\sqrt{N}(\hat{w} - w) \xrightarrow{d} W
\]

where \( W \) has a centered multivariate normal distribution on \( \mathbb{R}^d \).

In particular, we can write \( W = d\phi_w \cdot \text{diag}(\lambda)^{-1/2}U \), where \( U \) consists of independent, zero-mean Gaussian processes \( U_1, \ldots, U_d \) with covariance functions

\[
\Gamma_i(r, s) = S_i(r)S_i(s) \int_0^{r-s} \frac{d\Lambda_i}{S_iG_i(1 - \Delta\Lambda_i)} \quad i = 1, \ldots, d
\]

where \( \Lambda_i \) denotes the cumulative hazard function corresponding to \( S_i = \prod (1 - d\Lambda_i), i = 1, \ldots, d \); the symbol \( \prod \) denotes the product integral. Here, a minus sign in a subscript indicates the left-continuous version of a function and \( \Delta\Lambda = \Lambda - \Lambda_- \) is the jump size function of \( \Lambda \). Note that the covariance matrix of \( W \) is singular; in particular, \( (d\phi_w \cdot \text{diag}(\lambda)^{-1/2}U)_{ij} = 0 \) for all \( i = 1, \ldots, d \). The other entries \((i \neq j)\) are distributed as follows:

\[
(d\phi_w \cdot \text{diag}(\lambda)^{-1/2}U)_{ij} = \frac{1}{\sqrt{\lambda_i}} \int U_i^d dS_j - \frac{1}{\sqrt{\lambda_j}} \int U_j^d dS_i \sim N \left( 0, \frac{1}{\lambda_i} \int \Gamma_i^{\pm \pm} dS_j dS_i + \frac{1}{\lambda_j} \int \Gamma_j^{\pm \pm} dS_i dS_i \right)
\]

Here, the double appearance of \( \pm \) signs means the average of all four combinations of left- and right-continuous versions in both arguments of a two-parameter function.

Let us now turn to the estimation of the nonparametric concordance effects \( p \). A matrix multiplication of \( \hat{w} \) with \( E_d \) from the left is basically the same as taking the mean with respect to the inner index \( j \). This immediately brings us to the first main result:

**Theorem 1.** Under the asymptotic regime (4) we have

\[
\sqrt{N}(\hat{p} - p) := \sqrt{N}E_d(\hat{w} - w) \xrightarrow{d} E_dW = \left( \frac{1}{d} \sum_{i=1}^d \frac{1}{\sqrt{\lambda_i}} \int U_i^d dS_i - \frac{1}{d} \int U_j^d dS_i \right)_{j=1}^d
\]

where \( E_dW \) has the variance–covariance matrix \( V \) with the following entries

\[
V_{ii} = \frac{1}{\lambda_i} \int \Gamma_i^{\pm \pm} \hat{d}\hat{s} \left( \hat{S} - \frac{2}{d} \hat{S} \right) + \frac{1}{d^2} \sum_{j=1}^d \frac{1}{\lambda_j} \int \Gamma_j^{\pm \pm} dS_i dS_i
\]

in the \( i \)th diagonal entry, \( i = 1, \ldots, d \), and

\[
V_{ij} = \frac{1}{d^2} \sum_{j=1}^d \frac{1}{\lambda_j} \int \Gamma_j^{\pm \pm} dS_j dS_j - \frac{1}{d^2} \frac{1}{\lambda_j} \int \Gamma_j^{\pm \pm} dS_i dS_i
\]

in the off-diagonal entries \((i, j), i \neq j\).

A more compact form of the matrix \( V \) is given in Appendix 1.
3 Choice of test statistic

In order to develop hypothesis tests based on the estimator \( \hat{p} \), we next need to find a consistent estimator \( \hat{V}_N \) for \( V \). A natural choice is to plug in estimators for all unknown quantities that are involved in \( V \). In particular, we use the Kaplan–Meier estimators for all survival functions and \( \hat{\lambda}_i(t) = \hat{S}_i(t) \hat{S}_i(t) n_i \int_0^t \left[ Y_i(1 - \Delta \hat{\lambda}_i) \right]^{-1} \, d\hat{\lambda}_i \) for each covariance function \( \Gamma_i \) where \( Y_i \) is the number at risk process and \( \hat{\lambda}_i \) is the Nelson–Aalen estimator of the cumulative hazard matrix in group \( i \). Note that if \( \Delta \hat{\lambda}_i(u) = 1 \), we also have \( \hat{S}_i(u) = 0 \) in which case we let \( \hat{\lambda}_i(s, t) = 0 \) if \( s \geq u \) or \( t \geq u \). We denote the resulting covariance matrix estimator by \( \hat{V}_N = (\hat{V}_{ij})_{i,j \in [d]} \).

Lemma 2. Under the asymptotic regime (4), we have the consistency \( \hat{V}_N \overset{p}{\to} V \).

All of the developed convergence results are now utilized to find the most natural test statistic. First, note that the asymptotic covariance matrix \( V \) is singular since \( \mathbb{1}_N \sqrt{N}(\hat{p} - p) \equiv 0 \), whence \( r(V) = d - 1 \) follows. Furthermore, it is not at all obvious whether the ranks of the Moore–Penrose inverse \( r((CV_N C)^+) \) converge in probability to the rank \( r((CVC)^+) \) for a compatible contrast matrix C. Hence, the Wald-type statistic \( \hat{p}^T C (CV_N C)^+ C \hat{p} \) is not suitable for testing \( H_0^p(C) : Cp = 0 \): Its asymptotic behaviour is unclear and, hence, there is no reasonable choice of critical values.

Instead, we utilize a statistic that does not rely on the uncertain convergence of ranks of generalized inverses. This leads us to the survival version of the so-called ANOVA-rank-type statistic
\[
F_N(T) = \frac{N}{\text{tr}(TV_N)} \hat{p}^T T \hat{p}
\]
where \( T = C (CC)^+ C \) is the unique projection matrix onto the column space of \( C \). Below we analyze both, its asymptotic behaviour under null hypotheses of the form \( H_0^p(C) : Cp = 0 \) and under the corresponding alternative hypotheses \( H_1^p(C) : Cp \neq 0 \).

Theorem 2. Assume the asymptotic regime (4) and that \( \text{tr}(TV) > 0 \).

a) Under \( H_0^p(C) \) and as \( N \to \infty \), we have \( F_N(T) \overset{d}{\to} W E_dTE_dW / \text{tr}(TV) \) which is non-degenerate and non-negative with \( E(\chi) = 1 \).

b) Under \( H_1^p(C) \) and as \( N \to \infty \), we have \( F_N(T) \overset{p}{\to} \infty \).

As the distribution of \( \chi \) depends on unknown quantities (cf. Theorem 1), the test statistic \( F_N(T) \) in equation (5) is no asymptotic pivot. To nevertheless obtain proper critical values which lead to asymptotically exact inference procedures, we next propose and study a resampling approach.

4 Inference via multiplier bootstrap

In this section, we apply suitably tailored multiplier bootstrap techniques in order to approximate the small sample distribution of \( F_N(T) \). To this end, we consider the situation under \( H_0^p(C) \) in which case we may expand
\[
F_N(T) = \frac{N}{\text{tr}(TV_N)} (\hat{p} - p)^T T (\hat{p} - p) = \frac{N}{\text{tr}(TV_N)} (d\phi_S \cdot (\hat{S} - S)) / E_d T E_d (d\phi_S \cdot (\hat{S} - S)) + o_p(1)
\]
where \( \hat{S} \) is the vectorial aggregation of all Kaplan–Meier estimators \( \hat{S}_i, \ldots, \hat{S}_d \). First, we replace the martingale residuals, that are attached to the Kaplan–Meier estimators, with independent centered random variables which have approximately the same variance. In particular, we replace \( \sqrt{N}(S_i - S_i) \) with
\[
\hat{S}(t) \cdot \sqrt{N} \sum_{k=1}^{n_i} G_{ik} \int_0^t \left[ (Y_i(u) - \Delta N_i(u)) Y_i(u) \right]^{-1/2} dN_{ik}(u)
\]
A similar wild bootstrap Greenwood-type correction has been developed for tied survival and competing risks data.\(^{45}\) Here we utilized the usual counting process notation\(^{21}\): \( N_{ik} \) indicates whether the event of interest already took place for individual \( k \) in group \( i \). The wild bootstrap multipliers \( G_{ik}, i = 1, \ldots, n_i, i = 1, \ldots, d \), are i.i.d. with zero mean and unit variance and also independent of the data. A similar multiplier resampling approach has been applied to Nelson–Aalen and Aalen–Johansen estimators in one- and two-sample problems.\(^{46,47}\)
In a next step toward the construction of a wild bootstrap statistic, we replace \( d\phi_S \) with \( d\phi_S \). Let us denote the thus obtained wild bootstrap version of \( \sqrt{N}d\phi_S \) by \( W_N \). Conditionally on the data, this \( d^2 \)-variate random vector is for large \( N \) approximately normally distributed and its limit distribution coincides with that of \( W \); see the proof of Theorem 3 below for details.

Finally, a wild bootstrap version \( F_N^*(T) \) of \( F_N(T) \) requires that we also use a consistent wild bootstrap-type estimator \( tr(TV_N^*) \) of \( tr(TV_N) \). It is found by replacing the estimators \( \Gamma_i \) with

\[
\Gamma_i^*=\left(\hat{S}_i(s), \hat{S}_i(t)\right)_{ni} \sum_{k=1}^{n_i} G_{ik} \int_0^{\infty} \frac{dN_{ik}}{(Y_i - \Delta N_i)Y_i}.
\]

Its conditional consistency has been argued for which \( E(G_{11}) < \infty \) is a sufficient condition. These wild bootstrap-type variance estimators also have the nice interpretation of optional variation processes of the wild bootstrapped Kaplan–Meier estimators. Hence, the resulting wild bootstrap version of \( F_N(C) \) is

\[
F_N^*(T) = \frac{1}{tr(TV_N^*)} W_N^* E_d T E_d W_N^*
\]

The following conditional central limit theorem ensures the consistency of this resampling approach.

**Theorem 3.** Assume \( E(G_{11}) < \infty \) and that the conditions of Theorem 2 hold. Conditionally on \((X_{ik}, \delta_{ik}), i = 1, \ldots, d, k = 1, \ldots, n_i\), we have for all underlying values of \( p \)

\[
F_N^*(T) \overset{d}{\to} \chi = W E_d T E_d W / tr(TV)
\]

in probability as \( N \to \infty \).

We would like to stress that the limit distribution coincides with that of \( F_N(T) \) under \( H_0^p(C) \). For the wild bootstrap version \( F_N^*(T) \), however, the convergence result holds under both, the null and the alternative hypothesis, i.e. its conditional distribution always approximates the correct null distribution of the test statistic.

We conclude the theoretical part of this article with a presentation of deduced inference procedures for the effect sizes \( p \). To this end, let \( c_{N,a}^* \) denote the \((1 - \alpha)\)-quantile, \( \alpha \in (0, 1) \), of the conditional distribution of \( F_N^*(T) \) given the data. In practice, this quantile is approximated via simulation by repeatedly generating sets of the wild bootstrap multipliers \( G_{ki} \).

**Corollary 1.** Under the assumptions of Theorem 3, the test

\[
\varphi_N = 1\{F_N(T) > c_{N,a}^*\}
\]

is asymptotically exact and consistent. That is, \( E(\varphi_N) \to \alpha \cdot 1_{H_0^p(C)} + 1_{H_0^p(C)} \) as \( N \to \infty \).

### 4.1 Confidence intervals and regions

The above findings can also be used to construct confidence intervals and regions for the unknown effects. In particular, Theorem 1 in combination with the delta-method yields that

\[
CI = CI_{N,p}(p_i) = \left[ g^{-1}(g(\hat{p}_i)) \pm \frac{z_{1-\alpha/2}}{\sqrt{N}} \sqrt{V_{ii}(\hat{p}_i)} \right]
\]

is an asymptotic \((1 - \alpha)\) confidence interval for \( p_i (i = 1, \ldots, d) \) for any choice of a monotone and differentiable function \( g \) with \( g(p_i) \neq 0 \). Typical choices are \( g(x) = x \) or the range-preserving transformations \( g(x) = \logit(x) \) or \( g(x) = \text{probit}(x) \). Moreover, the asymptotic standard normal quantile \( z_{1-\alpha/2} \) in equation (6) could also be replaced by an appropriate wild bootstrap-based quantile. To this end, define \( \hat{p}_i^* \) via \( \sqrt{N}(\hat{p}_i^* - \hat{p}_i) = (E_d W_N^*)_i \). Since \( \hat{p}_i^* \) is not restricted to values in \((0, 1)\), an application of the probit- or logit-transformation is not possible on the bootstrap side. Hence, an appropriate quantile for the confidence interval for \( p_i \) is the \((1 - \alpha)\) quantile of the
conditional distribution of 
\[ \sqrt{N} |\hat{p}_i^s - \bar{p}_i| / \hat{V}_{ii}^{1/2} = |E_d W_{N,i}^*| / \hat{V}_{ii}^{1/2} \]
given the data. Here, \( \hat{V}_{ii}^* \) denotes the \( i \)-th diagonal element of \( V_{N,i}^* \). Using instead the \((1 - \alpha)\) quantile of the conditional distribution of the maximum of these \( d \) random variables automatically yields a family of confidence intervals for \( p_1, \ldots, p_d \) with an asymptotic family-wise confidence level of \((1 - \alpha)\).

To additionally obtain simultaneous confidence ellipsoids for vectors of effect contrasts, let \( r \) be the number of columns of \( C \) and denote by \( c_1, \ldots, c_r \) its column vectors. The presentation of a simultaneous confidence region for the contrasts \( c_\ell p, \ell = 1, \ldots, r \), in Corollary 2 below will be done in an implicit manner.

**Corollary 2.** Under the assumptions of Theorem 3, an asymptotically exact \((1 - \alpha)\)-confidence ellipsoid for the contrasts \( c_\ell p, \ell = 1, \ldots, r \), is given by

\[
CE = CE_{N,1-\alpha}(C) = \left\{ v \in \mathbb{R}^r : (C\tilde{p} - v)(CC')^{-1}(C\tilde{p} - v) \leq \frac{\text{tr}(T\tilde{V})}{N}N^{\alpha/2} \right\}
\]

That is, \( P(Cp \in CE) \to 1 - \alpha \) as \( N \to \infty \).

### 5 Simulations

In this section, we assess the small sample properties of the test \( \phi_N \) as proposed in Corollary 1.

#### 5.1 Behaviour under null hypotheses

We first focus on its type-I-error control with respect to

- various kinds of contrast matrices
- and different censoring intensities.

##### 5.1.1 Design and sample sizes

For ease of presentation, we restrict ourselves to a design with \( d = 6 \) groups with different sample size layouts: we considered small samples in a balanced design with \( n_1 = (n_1, \ldots, n_6)' = (10, 10, 10, 10, 10, 10)' \) and two unbalanced designs with \( n_2 = (n_1, \ldots, n_6)' = (10, 12, 14, 10, 12, 14)' \) and \( n_3 = (10, 12, 14, 14, 10, 12)' \), respectively. To obtain designs with moderate to large sample sizes, we increase these vectors component-wise by the factors \( K \in \{2, 3, 5, 10\} \). Moreover, depending on the question of interest, we below distinguish between a one-way layout with six independent groups and a \( 2 \times 3 \) two-way design.

##### 5.1.2 Censoring framework

We considered exponentially distributed censoring random variables \( C_{ij} \sim \text{Exp} (\lambda_i) \) with the following vectors \( \lambda = (\lambda_1, \ldots, \lambda_6)' \) of rate parameters: \( \lambda_1 = 0.4 \cdot 1, \lambda_2 = 0.5 \cdot 1, \lambda_3 = 2/3 \cdot 1, \lambda_4 = (0.4, 0.5, 2/3, 0.4, 0.5, 2/3)' \), \( \lambda_5 = (0.4, 0.5, 2/3, 2/3, 0.5, 0.4)' \), where \( 1 \in \mathbb{R}^6 \) is the vector consisting of 1s only. Thus, the first three settings correspond to equal censoring mechanisms with increased censoring rate from \( \lambda_1 \) to \( \lambda_3 \). The other two (\( \lambda_4 \) and \( \lambda_5 \)) lead to unequal censoring. By considering all 75 possible combinations, many possible effects of censoring and sample size assignments are analyzed. For example, in the set-up with \( n_2, K = 10 \) and \( \lambda_4 \), larger sample sizes are matched with a stronger censoring rate in an unbalanced design.

##### 5.1.3 Contrast matrices and null hypotheses

We simulated the true significance level of the tests for the null hypotheses \( H_{0i}(C) : Cp = 0 \) for two designs and different contrast matrices of interest:

In case of a one-way design with \( d = 6 \) groups, we were interested in the null hypotheses of ‘no group effect’ or ‘equality of all treatment effects’ \( H_{0i}(C) : (C_1p = 0) \). This may be described by considering the matrix \( C_1 = P_6 \), where here and below \( P_d = I_d - J_d/d \equiv I_d - 1d'1_d/d \) denotes the \( d \)-dimensional centering matrix.

Next, we consider a \( 2 \times 3 \) two-way layout with two factors \( A \) (with two levels) and \( B \) (with three levels). This is incorporated in Model (1) by setting \( d = a = 2 \cdot 3 = 6 \) and splitting up the index \( i \) into two indices \( i_1 = 1, 2 \) (for the levels of factor \( A \)) and \( i_2 = 1, 2, 3 \) (for the levels of factor \( B \)). Thus, we obtain survival times \( T_{i_1i_2k}, k = 1, \ldots, n_{i_1i_2} \).
and corresponding nonparametric concordance effects $p_{i_1i_2}$. More complex factorial designs can be incorporated similarly. In this $2 \times 3$ set-up, we are now interested in testing the null hypotheses of

(A) ‘No main effect of factor A’: $H_0^A(C_{2,A}) : \{C_{2,A} \mathbf{p} = \mathbf{0}\} = \{\bar{p}_1 = \bar{p}_2\}$,

(B) ‘No main effect of factor B’: $H_0^B(C_{2,B}) : \{C_{2,B} \mathbf{p} = \mathbf{0}\} = \{\bar{p}_1 = \bar{p}_2 = \bar{p}_3\}$ and

(AB) ‘No A \times B interaction effect’: $H_0^{AB}(C_{2,AB}) : \{C_{2,AB} \mathbf{p} = \mathbf{0}\} = \{p_{i_1i_2} = \bar{p}_{i_1} - \bar{p}_{i_2} = \bar{p}_s = 0\}$ for all $i_1, i_2$

where $\bar{p}_{i_1}, \bar{p}_{i_2}$, and $\bar{p}_s$ denote the means over the dotted indices. In particular, the corresponding contrast matrices are given by $C_{2,A} = \mathbf{P}_2 \otimes \frac{1}{2} \mathbf{J}_3$, $C_{2,B} = \frac{1}{2} \mathbf{J}_2 \otimes \mathbf{P}_3$, and $C_{2,AB} = \mathbf{P}_2 \otimes \mathbf{P}_3$, where $\otimes$ indicates the Kronecker product.

5.1.4 Survival distributions

For ease of presentation, we only considered a rather challenging scenario, where the groups follow different survival distributions. In particular, we simulated

(G1) a lognormal distribution with meanlog parameter 0 and sdlog parameter 0.2726 for the first group,
(G2) a Weibull distribution with scale parameter 1.412 and shape parameter 1.1 for the second group,
(G3) a Gamma-distribution with scale parameter 0.4 and shape parameter 2.851 for the third group and
(G4–G6) mixing distributions of all pair combinations of the first three survival functions for the last three groups.

The first three survival functions are illustrated in Figure 1. We note that preliminary simulations for more crude scenarios with identical survival distributions in all groups exhibited a much better type-I-error control of our testing procedure (results not shown). Anyhow, the parameters of the above distributions were chosen in such a way that the nonparametric concordance effects of all groups are equal, i.e. $p_i = 0.5$ for all $i = 1, \ldots, 6$ (one-way) and $p_{i_1i_2} = 0.5$ for all $i_1 = 1, 2; i_2 = 1, 2, 3$ (two-way), respectively. Thus, all considered null hypotheses are true. We would like to stress that the case of continuously distributed survival times corresponds to an infinite-dimensional problem and is thus more difficult than the discrete case. For example, this observation has been confirmed in another simulation study\textsuperscript{44}; the convergence rate of the actual coverage probabilities of confidence bands to the nominal confidence level is much faster the more discretely distributed the survival data is. Moreover, to make the simulation scenario even more challenging, we considered the situation with infinite $r$ to also get an indication of the functionality of the test in this case.

5.1.5 Simulations

We chose as wild bootstrap multipliers centered unit Poisson variables because a formal Edgeworth expansion and two simulation studies have indicated that those have theoretical and practical advantages over the common choice of standard normal multipliers.\textsuperscript{48} We chose the nominal level $\alpha = 5\%$ and conducted each test 10,000

![Figure 1. Survival functions underlying the first three simulated sample groups.](image)
times for $K=1, 2$, and $3$ and $5000$ times for $K=5$ and $10$ because of the massively increasing computational complexity for large samples. Each test was based on critical values that were found using 1999 wild bootstrap iterations. All simulations were conducted with the help of the R computing environment.49

5.1.6 Results

The true type-I-error results for the four different null hypotheses are shown in Table 1 (upper panel: one-way for $H_0^0(C_1)$ and lower panel: two-way for $H_0^0(C_{2,1})$ and Table 2 (two-way for $H_0^0(C_{2,1})$ in the upper and $H_0^0(C_{2,AB})$ in the lower panel). It is apparent that all simulated levels are elevated for the smallest sample sizes ($K=1$), especially for the one-way test: here almost all type-I-error probabilities are between $13.0\%$ and $17.7\%$. For the two-way tests, these probabilities are mainly between $8.1\%$ and $11.7\%$ in this case ($K=1$). On the other hand, this is due to the relatively strong censoring rates: for $\lambda = 0.4$, the censoring probabilities across all sample groups are between $33\%$ and $37\%$ (found by simulating 100,000 censoring and survival time random variables each); for $\lambda = 0.5$, these probabilities range from $39.5\%$ to $41.5\%$; and for $\lambda = 2/3$, they even reach values between $48.5\%$ and $49\%$; resulting in only $5$ to $7$ uncensored observations per group. On the other hand, not to restrict the time horizon in inferential procedures about survival functions appears to slightly slow down the convergence of type-I-error probabilities to the nominal level as the sample size increases; similar findings have been obtained in the context of confidence bands for unrestricted survival functions.50 However, the error probabilities recover for samples of double size (i.e. between 20 and 28) already: in the one-way design, these error rates drop to mainly $8.2\%–9.9\%$, and in all two-way tests, we even achieve rates of mainly $6.1\%–8\%$. If the sample sizes are tripled (i.e. between 30 and 42), most of the type-I-error probabilities are between $7\%$ and $8\%$ (one-way) or $5.2\%$ and $6.9\%$ (two-way). In case of

| $n$ | $\lambda/K$ | 1   | 2   | 3   | 5   | 10  |
|-----|-------------|-----|-----|-----|-----|-----|
| $\lambda_1$ | $\lambda_1$ | 14.7 | 8.8 | 7.2 | 6.4 | 5.7 |
|     | $\lambda_2$ | 16.6 | 9.3 | 7.7 | 6.3 | 5.7 |
|     | $\lambda_3$ | 19.7 | 11.0 | 8.6 | 6.6 | 5.8 |
|     | $\lambda_4$ | 17.7 | 9.9 | 7.7 | 6.1 | 5.9 |
|     | $\lambda_5$ | 17.4 | 9.5 | 7.8 | 6.7 | 6.3 |
| $\lambda_1$ | $\lambda_1$ | 13.0 | 7.9 | 6.7 | 5.9 | 5.7 |
|     | $\lambda_2$ | 13.9 | 8.4 | 7.1 | 6.9 | 5.4 |
|     | $\lambda_3$ | 16.5 | 9.1 | 7.8 | 6.2 | 5.8 |
|     | $\lambda_4$ | 14.9 | 8.5 | 7.3 | 6.0 | 5.6 |
|     | $\lambda_5$ | 14.6 | 9.0 | 6.8 | 6.3 | 5.6 |
| $\lambda_1$ | $\lambda_1$ | 12.2 | 8.2 | 7.1 | 6.0 | 5.2 |
|     | $\lambda_2$ | 13.7 | 8.6 | 7.0 | 6.1 | 5.3 |
|     | $\lambda_3$ | 17.7 | 9.3 | 7.6 | 6.8 | 5.9 |
|     | $\lambda_4$ | 14.8 | 8.8 | 7.7 | 6.3 | 5.9 |
|     | $\lambda_5$ | 14.1 | 8.6 | 7.2 | 6.3 | 5.9 |
| $\lambda_1$ | $\lambda_1$ | 9.0 | 6.3 | 6.0 | 5.7 | 5.4 |
|     | $\lambda_2$ | 9.0 | 6.7 | 6.0 | 5.6 | 4.6 |
|     | $\lambda_3$ | 10.5 | 7.0 | 6.3 | 5.9 | 5.6 |
|     | $\lambda_4$ | 10.1 | 6.7 | 6.1 | 5.7 | 5.7 |
|     | $\lambda_5$ | 9.4 | 6.2 | 5.8 | 5.6 | 5.0 |
| $\lambda_1$ | $\lambda_1$ | 7.7 | 6.3 | 5.8 | 5.3 | 5.9 |
|     | $\lambda_2$ | 8.1 | 6.3 | 6.0 | 6.0 | 5.3 |
|     | $\lambda_3$ | 9.7 | 6.8 | 6.4 | 5.0 | 5.4 |
|     | $\lambda_4$ | 8.2 | 6.3 | 6.2 | 5.9 | 5.4 |
|     | $\lambda_5$ | 8.9 | 6.7 | 6.2 | 5.4 | 5.1 |
| $\lambda_1$ | $\lambda_1$ | 7.8 | 6.3 | 6.0 | 5.6 | 4.9 |
|     | $\lambda_2$ | 8.5 | 6.2 | 5.5 | 5.3 | 5.0 |
|     | $\lambda_3$ | 9.2 | 6.9 | 6.5 | 5.6 | 5.1 |
|     | $\lambda_4$ | 8.4 | 6.1 | 6.1 | 5.8 | 4.5 |
|     | $\lambda_5$ | 8.2 | 6.7 | 5.7 | 6.0 | 5.7 |
the sample size factor $K = 5$, all results are only slightly liberal, and for $K = 10$ (i.e. sample sizes between 100 and 140), we see that the nominal level is well attained.

### 5.2 Behaviour under shift alternatives

In addition to the simulations of the previous subsection, we also conducted a small power simulation of the above tests. For the alternative hypotheses, we considered a shift model: taking the same six basic survival and censoring functions as in the first set of simulations, we shift all survival and censoring times of the first sample group by $d \in \{0.1, 0.2, \ldots, 1\}$. In this way, we maintain the same censoring rates as before and the distance to the null hypotheses is gradually increased: for growing $d > 0$, we obtain a growing relative effect $p_{1} > 0.5$ (one-way) and $p_{11} > 0.5$ (two-way), respectively. For each of the above considered contract matrices, $C_1, C_{2,A}, C_{2,B}, C_{2,AB}$, we conducted one set of simulations with different unbalanced sample sizes and censoring rate combinations. For each set-up, we increased the sample sizes by the factors $K = 1, 3, 5$. The results are displayed in Figure 2.

We see that, even for the smallest sample sizes (between 10 and 14), the power of the two-way testing procedures increase to 0.5 or 0.6 as the shift parameter approaches 1. For larger samples sizes, the theoretically proven consistency is apparent. In comparison, the one-way test has a much higher power: For the undersized case ($K=1$), it already reaches a power of 0.8 while for moderate to larger sample sizes the power is almost 1 for shift parameters $d \geq 0.5$. In comparison to the two-way procedure, its superior power is, however, partially paid at the price of its pronounced liberality; especially for small sample sizes.

All in all, the simulations confirm that all tests have a satisfactory power with increasing sample size and/or shift parameter while maintaining a reasonable control of the nominal level for sample sizes of 30 to 42 already.

| $n$ | $\lambda/K$ | 1   | 2   | 3   | 5   | 10  |
|-----|-------------|-----|-----|-----|-----|-----|
| $n_1$ | $\lambda_1$ | 10.0| 7.2 | 6.4 | 6.2 | 6.0 |
|      | $\lambda_2$ | 11.4| 7.7 | 6.7 | 5.9 | 4.9 |
|      | $\lambda_3$ | 13.4| 8.0 | 6.9 | 5.9 | 5.9 |
|      | $\lambda_4$ | 12.2| 7.6 | 6.9 | 6.1 | 5.9 |
|      | $\lambda_5$ | 12.1| 7.5 | 6.7 | 5.9 | 5.9 |
| $n_2$ | $\lambda_1$ | 9.5 | 6.6 | 6.1 | 6.0 | 5.0 |
|      | $\lambda_2$ | 10.2| 7.4 | 6.5 | 6.0 | 5.5 |
|      | $\lambda_3$ | 11.6| 7.8 | 6.6 | 5.6 | 5.7 |
|      | $\lambda_4$ | 10.4| 7.0 | 6.4 | 5.5 | 6.1 |
|      | $\lambda_5$ | 10.2| 7.1 | 6.2 | 5.9 | 5.4 |
| $n_3$ | $\lambda_1$ | 9.5 | 7.2 | 5.8 | 5.2 | 5.2 |
|      | $\lambda_2$ | 9.6 | 6.8 | 6.3 | 5.4 | 5.6 |
|      | $\lambda_3$ | 11.6| 7.4 | 6.9 | 6.2 | 5.6 |
|      | $\lambda_4$ | 9.9 | 7.4 | 6.2 | 6.5 | 5.4 |
|      | $\lambda_5$ | 9.7 | 7.2 | 6.4 | 5.7 | 5.0 |

| $n$ | $\lambda/K$ | 1   | 2   | 3   | 5   | 10  |
|-----|-------------|-----|-----|-----|-----|-----|
| $n_1$ | $\lambda_1$ | 10.1| 7.2 | 6.3 | 5.7 | 5.3 |
|      | $\lambda_2$ | 11.2| 7.2 | 6.2 | 5.9 | 5.1 |
|      | $\lambda_3$ | 13.3| 8.5 | 7.0 | 6.5 | 5.5 |
|      | $\lambda_4$ | 11.6| 7.8 | 6.6 | 6.1 | 5.3 |
|      | $\lambda_5$ | 11.6| 7.7 | 6.4 | 5.9 | 5.6 |
| $n_2$ | $\lambda_1$ | 9.2 | 6.6 | 5.9 | 6.2 | 5.3 |
|      | $\lambda_2$ | 9.8 | 6.9 | 6.6 | 5.4 | 5.7 |
|      | $\lambda_3$ | 11.7| 7.5 | 6.4 | 5.8 | 5.4 |
|      | $\lambda_4$ | 9.8 | 6.8 | 6.4 | 5.2 | 5.0 |
|      | $\lambda_5$ | 10.8| 7.1 | 6.2 | 5.8 | 5.6 |
| $n_3$ | $\lambda_1$ | 8.3 | 6.6 | 6.3 | 5.2 | 5.1 |
|      | $\lambda_2$ | 9.6 | 6.9 | 5.9 | 5.8 | 5.7 |
|      | $\lambda_3$ | 11.2| 7.6 | 6.4 | 5.3 | 5.8 |
|      | $\lambda_4$ | 10.4| 6.9 | 5.8 | 5.4 | 4.7 |
|      | $\lambda_5$ | 9.8 | 6.8 | 5.5 | 5.5 | 5.7 |
6 Data example

We illustrate the developed theory on a dataset from a colon cancer study. Considering the patients in Stage C, that is, there had been metastases to regional lymph nodes, the data consist of eligible 929 patients suffering from colon cancer. Survival (measured in days) was the primary endpoint of the study. We focus on the two factors ‘gender’ and ‘treatment’ (with three levels) to obtain a crossed $2 \times 3$ survival design which is in line with a setting from our simulation study. In particular, there were 315 patients in the observation group, 310 others were treated with levamisole, and 304 received levamisole, combined with fluorouracil. Levamisole was originally used as an anthelmintic drug and fluorouracil (5-FU) is a medicine to treat various types of cancer. The patients in the study...
had been randomized into one of these three treatment groups. Also, there were nearly as many women (445) as men (484) involved in the study. Figure 3 depicts the Kaplan–Meier estimates of the survival probabilities for each treatment × sex subgroup. We refer to the article by Moertel et al. for more details about the study. The dataset is freely accessible via the R command data(colonCS) after having loaded the package condSURV. The aim is now to investigate the presence of main or interaction effects of treatment and gender. As there are several ties in the data (roughly 16%; see the supplement for details) and we do not want to impose specific distributional assumptions, we focus on the nonparametric concordance effects. To this end, we first have to choose a proper τ. From our retrospective view, the most reasonable choice is found by determining for each group the minimal observed censoring time that exceeds all observed survival times in that group. We call these censoring times “terminal times”. Then, τ is set to be the minimal terminal time. In doing so, the group with that minimum terminal time does not benefit nor does it suffer from having the earliest terminal time when compared to the other groups.

The first block in Table 3 shows the sample sizes of all subgroups. In the present data example, the minimal terminal time is τ = 2173; see the second block of Table 3. In view of the sample sizes and the censoring rates given in the third block of Table 3, we compare the present dataset with the simulation set-ups in the previous section: a similarly strong censorship is obtained for 𝜆3 and comparable sample sizes n ∈ [100, 140] for the choice K = 10. Thus, judging from the rightmost columns of Tables 1 and 2, we find it is safe to assume actual type-I-error probabilities of about 5.1–5.9% of the proposed nonparametric one- and two-way survival tests.

We tested the data in one- and two-factorial set-ups and chose 𝛼 = 5% as the significance level. As in the simulation study, we used B = 1999 bootstrap iterations for each test. For the tests in the two-factorial model, we considered the null hypotheses corresponding to no main treatment effect, no main effect in sex, and no interaction effect between both. The test results, by means of p-values, are shown in Table 4. It should be noted that the p-values have not been adjusted for a type-I-error multiplicity but the test decisions remain the same after an application of, say, the simple Bonferroni procedure.

We found a significant indication against the equality of all d = 6 groups (p-value < 0.001). Judging from the main effects, this overall group difference may be explained by the significant treatment effect (p-value < 0.001) while no difference between the sexes (p-value = 0.331) were found. However, a significant interaction effect between treatment and sex (p-value < 0.001) make these conclusions uncertain. This is why we split the data by the factor sex and repeat the analysis for the female and male group within two separate 1-factorial 1 × 3 frameworks: Our hypothesis tests yielded p-values of < 0.001 for males and 0.49 for females, indicating a treatment effect for male patients only. Indeed, looking at the rightmost block of Table 3, we agree with the

| Treatment                  | Sample size | Terminal time | Censoring rate | Effect size |
|----------------------------|-------------|---------------|----------------|-------------|
|                            | Male        | Female        | Male           | Female      | Male | Female | Male | Female |
| Observation                | 166         | 149           | 2800           | 2562        | 47.6 | 51.0   | 0.475| 0.483  |
| Levamisole                 | 177         | 133           | 2915           | 2173        | 47.5 | 52.6   | 0.459| 0.501  |
| Levamisole plus fluorouracil| 141        | 163           | 2726           | 2198        | 68.8 | 55.2   | 0.581| 0.501  |

Note: Columns: sex, row: treatments.

| Null hypothesis                        | H0 ( ) | Set-up      | p-value |
|----------------------------------------|--------|-------------|---------|
| Equality of all effects                | H0 (C1) | One-factorial | <0.001  |
| No main effect in sex                  | H0 (C2A) | two-factorial | 0.331   |
| No main effect in treatment            | H0 (C2,8) | two-factorial | <0.001  |
| No interaction effect                  | H0 (C2,AB) | two-factorial | <0.001  |
findings of the hypothesis tests: the gender effect seems to be canceled out if the treatment groups are combined, but within the male gender there seems to be a big difference in the concordance effects \((p_{11} \in [0.459, 0.581])\). Also, the interaction effect is apparent, as the female groups do not seem to strongly benefit from any treatment \((p_{23} \in [0.483, 0.501])\). On the other hand, the male groups exhibit a worse than average survival probability in the observation and the levamisole treatment group \((p_{12} = 0.475, p_{12} = 0.459)\) but a much better than average survival probability for the combination treatment \((p_{13} = 0.581)\). Here the value \(p_{13} = 0.581\) roughly means that a randomly generated observation from this specific group survives a randomly generated observation from the mean distribution of all groups with probability 58.1%. We would like to note that the values of \(p_{11}, \ldots, p_{23}\) changed only marginally when we altered the value of \(\tau\) to 1500, 2600, or 3000. Also, the conclusions drawn from the hypothesis tests remain the same, even though the theory developed in this paper does not apply to the too large choices of \(\tau \in \{2600, 3000\}\); see the supplement for details.

Taking another look at the Kaplan–Meier curves in Figure 3, we immediately see that our concordance effects and the test outcomes make sense. We clearly see that there is a big difference in the male survival probabilities (the combination treatment group is superior to the levamisole treatment group which is in turn superior to the observation group) but there is not much of a difference between the female groups’ survival curves. Indeed, comparing the Kaplan–Meier curve of the pooled males’ survival times with that of the pooled females’ times, we graphically find no evident main gender effect. The plot of both Kaplan–Meier estimators is shown in the supplement.

As suggested by a referee, we compared the results of our nonparametric factorial analysis to the outcomes of a similar Cox model analysis while adjusting for ties in the event times. The `coxph`-method in the R-package `survival` offers the following three choices for adjusting for ties:\(^{53}\)

- `efron` is the default `coxph` choice. It is quite accurate in case of a large number of ties.
- `breslow` is rather easy to implement and the default choice of many other program packages.
- `exact` is based on logistic regression and it is appropriate for a discrete time scale and a small number of unique event times. Furthermore, it is computationally quite expensive.

Another option is to break the ties by adding small noise to the event times. This approach is not applied to the present data example because we explicitly wish to compare methods for tied survival times. In a first typical Cox analysis, we included `sex` and `treatment` as additive factor effects, as well as their interaction. The baseline hazard is that of a female patient in the observation subgroup. Table 5 presents the results obtained from `summary` applied to the `coxph` command with the `ties=` `efron` option to handle tied event times: point estimates of the parameters, \(p\)-values, and 95% confidence intervals based on the asymptotic normality of the point estimates. The other two methods for handling ties yield essentially the same results.

As explained above, the results need to be interpreted in comparison to the female observation group. We again see that the interaction effect of male gender and levamisole plus fluorouracil is significantly beneficial at level \(\alpha = 5\%\). Also, the remaining point estimates of the parameters point towards the same direction as in the nonparametric analysis, but all of them are non-significant. Moreover, adjusting for the type-I-error multiplicity (by Bonferroni or Holm), no effect was detected at a significance level of 5%.

Our expert referee also suggested to perform a Cox analysis that matches our nonparametric one. To this end, we separately tested the presence of an interaction effect in the Cox model by using the parameter estimates related to the covariates `sex` \(\times\) `levamisole` and `sex` \(\times\) `levamisole` plus fluorouracil. Denoting the unknown interaction effect parameters by \(\theta_1\) and \(\theta_2\), we used the two-degree-of-freedom Wald test for testing \(H_0^\theta: \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \) against

| Covariate                         | Estimate | \(p\)-Value | 95% CI     |
|-----------------------------------|----------|-------------|------------|
| Sex                               | 0.074    | 0.634       | \([-0.230, 0.377]\) |
| Levamisole                        | -0.128   | 0.451       | \([-0.461, 0.205]\) |
| Levamisole plus fluorouracil      | -0.149   | 0.357       | \([-0.467, 0.169]\) |
| Sex \(\times\) levamisole        | 0.169    | 0.450       | \([-0.270, 0.608]\) |
| sex \(\times\) levamisole plus fluorouracil | -0.489 | 0.043       | \([-0.962, -0.016]\) |
Irrespective of the particular method of tie-adjustment, the Wald test yielded a $p$-value of approximately 0.023. Hence, at level $\alpha = 5\%$, this again confirms the presence of a sex-treatment interaction effect, i.e. also in the Cox model. We would like to note that this test decision is independent of the particular choice of the baseline hazard function. This significant interaction effect again motivates two separate analyses: first, we fitted a Cox model for the male subgroup and both treatment-related dummy covariates. The two-degree-of-freedom Wald test on the involved parameters yielded a highly significant result ($p < 0.001$). Hence, also the Cox model-based male subgroup analysis confirms the presence of a treatment effect. A similar analysis for the female subgroup again resulted in a non-significant $p$-value $\approx 0.62$. We note that this course of action generally only controls the family wise error rate if the hypotheses would have been formulated as a hierarchically testing problem beforehand. Otherwise, a simple Bonferroni correction would alter the results as the interaction effect would no longer be significant. However, a more enhanced adjustment by Holm’s method would also lead to the conclusion of the hierarchical testing problem.

### 6.1 Comparison to the nonparametric analysis

The confidence intervals which resulted from the Cox analysis have to be interpreted on the log-hazard ratio level. This is much more complicated than on the level of probabilities of survival superiority, as in our nonparametric analysis. Also, these outcomes of the Cox analysis depend on the particular choice of the baseline hazard, which in this case refers to females from the observation group. Only the Wald tests for any interaction effect and for the treatment effect in the subgroup analyses are independent of the baseline reference subgroup. Our nonparametric method, on the other hand, always compares the single groups to the all-group average, which appears more natural in the present example.

To further demonstrate the good interpretability of the nonparametric analysis, we also derived simultaneous 95% confidence intervals for all nonparametric concordance effects; see Table 6.

| Treatment            | Sex   | Multiplier bootstrap | Normal approximation |
|----------------------|-------|-----------------------|----------------------|
| Observation          | male  | [0.447, 0.502]        | [0.446, 0.504]       |
|                      | female| [0.454, 0.512]        | [0.452, 0.514]       |
| Levamisole           | male  | [0.432, 0.487]        | [0.430, 0.489]       |
|                      | female| [0.471, 0.531]        | [0.469, 0.532]       |
| Levamisole plus fluorouracil | male  | [0.553, 0.608]        | [0.551, 0.609]       |
|                      | female| [0.473, 0.530]        | [0.471, 0.532]       |

Finally, we relate our results to the original findings of analyses that involve the Cox proportional hazards model and logrank tests. It has been detected that “Therapy with levamisole plus fluorouracil produced an unequivocal advantage over observation” and that levamisole alone did not produce a detectable effect. Furthermore, they concluded from an exploratory subset analysis that the “levamisole-fluorouracil treatment appeared to have the greatest advantage among male patients [...].” This is exactly what we confirm in our nonparametric analyses. However, Moertel et al. apparently did not account for the ties which are present in the data nor did they clearly stress the rather weak effect of the levamisole-fluorouracil treatment for women. They just state that their “results show [...] striking contradictions to those of subset analyses reported in the NCCTG study, in which levamisole plus fluorouracil was found to be most effective in reducing the risk of recurrence among female patients [...].” among other subgroups of patients.
7 Discussion

We proposed novel nonparametric inference procedures for the analysis of factorial survival data that may be subject to independent random right-censoring. Critical values are obtained from a multiplier wild bootstrap approach which led to asymptotically valid tests and confidence regions for meaningful effect parameters. Thereby, the procedures do not require any multiplicative or additive hazard structure nor specific distributional survival and censoring assumptions. In particular, different group distributions are allowed and ties are accounted for accordingly. Moreover, different to other nonparametric survival procedures, our methods are not only driven towards hypothesis testing but also to uncertainty quantification of the underlying effect estimators. The latter can be used to comprehensibly describe and infer main and interaction effects in general nonparametric factorial survival designs with an arbitrary number of fixed factors. Together with a 1–1 connection with hazard ratios in proportional two-sample designs and the possibility to construct simultaneous confidence intervals for all nonparametric effects of interest, this makes the new methods appealing for practical purposes.

To investigate their theoretical properties, we rigorously proved central limit theorems of the underlying statistics and consistency of the corresponding procedures. In addition, extensive simulations were conducted for one- and two-way designs to also assess their finite sample properties in terms of power and type-I-error control. In case of small sample sizes with less than 10 completely observed subjects per group, they revealed a liberal behaviour; especially for the one-way testing procedure. However, for moderate to larger sample sizes, the asymptotic results kicked in and the stated theoretical results were recovered.

Finally, the methods were used to exemplify the analysis of survival data in a study about treatments for colon cancer patient within a two-factorial survival design. As severe ties were present in the data, classical hazard-based methods were not directly applicable and adjustments for ties were required. In comparison, our newly proposed nonparametric methods provided a very decent alternative for the analysis of such factorial survival designs without postulating any strict assumptions.

To allow for a straightforward application, it is planned to implement the procedure into an easy to use R-package. In future research, we will consider the case of stochastically ordered subgroups, for which a multiple testing algorithm could be developed with the aim to detect significantly different collections of all subgroups: subgroups with no significant differences in the nonparametric concordance effects may be combined to facilitate the interpretation of the outcomes and to ultimately serve for the development of different, more personalized medicines, one for each new subgroup combination. Moreover, extensions of the current methodology to ordered alternatives or factorial designs obtained via stratified sampling will be part of a practically useful consecutive testing procedure. Another future research topic covers the question on how to develop comparable procedures for causal inference in general factorial finite population models. Here, the non-identifiability problem of effect measures has to be resolved accordingly.

Acknowledgements

Both authors would like to thank two referees and the editor whose help has substantially improved the manuscript.

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: Markus Pauly wishes to thank for the support from the German Research Foundation (Deutsche Forschungsgemeinschaft).

ORCID iD

Dennis Dobler http://orcid.org/0000-0002-9040-0854

Supplementary Material

Supplemental material for this article is available online.
References

1. Cassidy J, Clarke S, Díaz-Rubio E, et al. Randomized phase III study of capecitabine plus oxaliplatin compared with fluorouracil/folinic acid plus oxaliplatin as first-line therapy for metastatic colorectal cancer. J Clin Oncol 2008; 26: 2006–2012.

2. The International Study Group. In-hospital mortality and clinical course of 20891 patients with suspected acute myocardial infarction randomised between alteplase and streptokinase with or without heparin. Lancet 1990; 336: 71–75.

3. Baigent C, Collins R, Appleby P, et al. ISIS-2: 10 year survival among patients with suspected acute myocardial infarction in randomised comparison of intravenous streptokinase, oral aspirin, both, or neither. BMJ 1998; 316: 1337.

4. Kurz A, Fleischmann E, Sessler DI, et al. Effects of supplemental oxygen and dexamethasone on surgical site infection: a factorial randomized trial. Brit J Anaesth 2015; 115: 434–443.

5. Mehta SR, Tanguay JF, Eikelboom JW, et al. Double-dose versus standard-dose clopidogrel and high-dose versus low-dose aspirin in individuals undergoing percutaneous coronary intervention for acute coronary syndromes (CURRENT-OASIS 7): a randomised factorial trial. Lancet 2010; 376: 1233–1243.

6. Lubsen J and Pocock SJ. Factorial trials in cardiology: pros and cons. Eur Heart J 1994; 15: 585–588.

7. Brunner E and Dette H and Munk A. Box-type approximations in nonparametric factorial designs. J Am Stat Assoc 1997; 92: 1494–1502.

8. Brunner E and Puri ML. Nonparametric methods in factorial designs. Stat Pap 2001; 42: 1–52.

9. Gao X and Alvo M. A unified nonparametric approach for unbalanced factorial designs. J Am Stat Assoc 2005; 100: 926–941.

10. Gao X and Alvo M. Nonparametric multiple comparison procedures for unbalanced two-way layouts. J Stat Plan Infer 2008; 138: 3674–3686.

11. Gao X, Alvo M, Chen J, et al. Nonparametric multiple comparison procedures for unbalanced one-way factorial designs. J Stat Plan Infer 2008; 138: 2574–2591.

12. Akritas M.G. Nonparametric Models for ANOVA and ANCOVA Designs. In: Lovric M. (eds) International Encyclopedia of Statistical Science. Berlin, Heidelberg: Springer, 2011, pp. 964–968.

13. Dutta S and Datta S. A rank-sum test for clustered data when the number of subjects in a group within a cluster is informative. Biometrics 2016; 72: 432–440.

14. Friedrich S, Konietzchke F and Pauly M. A wild bootstrap approach for nonparametric repeated measurements. Comput Stat Data Anal 2017; 113: 38–52.

15. Umlaßt M, Konietzchke F and Pauly M. Rank permutation approaches for nonparametric factorial designs. Brit J Math Stat Psychol 2017; 70: 368–390.

16. Konietzchke F, Friedrich S, Brunner E, et al. rankFD: Rank-based tests for general factorial designs, 2016, https://CRAN.R-project.org/package=rankFD. R package v. 0.0.1.

17. Brunner E, Konietzchke F, Pauly M, et al. Rank-based procedures in factorial designs: hypotheses about non-parametric treatment effects. J Roy Stat Soc B Methodol 2017; 79: 1463–1485.

18. Strobl M, Placzek M, Konietzchke F, et al. Wild bootstrapping rank-based procedures: multiple testing in nonparametric split-plot designs. J Multivariate Anal 2019; 171: 176–192.

19. Dobler D, Friedrich S and Pauly M. Nonparametric MANOVA in Mann–Whitney effects. arXiv preprint arXiv:1712.06983v2, 2018..

20. Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother Rep 1966; 50: 163–170.

21. Andersen PK, Borgan Ø, Gill RD, et al. Statistical models based on counting processes. New York, NY: Springer, 1993.

22. Ehm W, Mammen E and Müller DW. Power robustification of approximately linear tests. J Am Stat Assoc 1994; 89: 1025–1033.

23. Liu PY and Dahlberg S. Design and analysis of multiarm clinical trials with survival endpoints. Control Clin Trials 1995; 16: 119–130.

24. Janssen A and Neuhaus G. Two-sample rank tests forensored data with non-predictable weights. J Stat Plan Infer 1997; 60: 45–59.

25. Bathke A, Kim MO and Zhou M. Combined multiple testing by censored empirical likelihood. J Stat Plan Infer 2009; 139: 814–827.

26. Yang S and Prentice R. Improved logrank-type tests for survival data using adaptive weights. Biometrics 2010; 66: 30–38.

27. Fleming TR and Harrington DP. Counting processes and survival analysis. John Wiley & Sons, 2011.

28. Brendel M, Janssen A, Mayer C-D, et al. Weighted Logrank permutation tests for randomly right censored life science data. Scand J Stat 2014; 41: 742–761.

29. Cox DR. Regression models and life-tables. J Roy Stat Soc B Met 1972; 34: 187–220.

30. Scheike TH and Zhang M-J. An additive–multiplicative Cox–Aalen regression model. Scand J Stat 2002; 29: 75–88.

31. Scheike TH and Zhang M-J. Extensions and applications of the Cox-Aalen survival model. Biometrics 2003; 59: 1036–1045.

32. Green S, Liu P-Y and O’Sullivan J. Factorial design considerations. J Clin Oncol 2002; 20: 3424–3430.
The proof of this lemma extends previous results from a univariate context.\textsuperscript{37} It has been shown that the functional $\phi : (BV[0, \tau])^2 \to \mathbb{R}, (f, g) \mapsto -\int_0^\tau f^g(u) \, du$ is Hadamard-differentiable in functions of bounded variation; see the proof of their Theorem 1 in the supplement to that article for more details. Hence, a similar result transfers to the functional $\phi$ in the present article because $\phi$ is simply a multivariate aggregation of functionals of the same type as $\phi$.

For the sake of completeness, we point out that the $(i - 1)d + j$th entry of continuous and linear Hadamard-derivative of $\phi$ at $S$ is given by

$$(d\phi_S \cdot h)_{ij} = \int h_i^x \, dS_j - \int h_j^x \, dS_i, \quad h = (h_1, \ldots, h_d) \in (\mathcal{D}[0, \tau])^d$$

if $h_1(0) = \cdots = h_d(0) = 0$ which indeed is the case in the present application.
1.2 Proof of Theorem 1

This convergence result follows from Lemma 1 in combination with the continuous mapping theorem.

1.3 Proof of Lemma 2

It is well known that each Kaplan–Meier estimator \( \hat{S} \), and also each covariance function estimator \( \hat{\Gamma} \), is uniformly consistent on \([0, \tau]\) for \( S \) and on \([0, \tau]^2\) for \( \Gamma \), respectively, \( i = 1, \ldots, d \). Note here that due to the assumption \( P(T_i > \tau) > 0 \) for all groups \( i \), the cumulative hazard functions are bounded on the interval of interest: \( \Lambda_i(\tau) < \infty \).

Since \( \hat{V}_N \) is a continuous functional of these estimators, it follows from the continuous mapping theorem that the matrix \( \hat{V}_N \) is consistent for \( V \). Likewise, we conclude that the estimator \( tr(T \hat{V}_N) \) is consistent for the trace \( tr(TV) \) due to continuity of the trace and linear maps.

1.4 Proof of Theorem 2

Again, the result essentially follows from the continuous mapping theorem as \( p \mapsto p' T p \) is a continuous map. Thus, taking Lemma 2 into account, an application of Slutsky’s lemma reveals that \( F_N(C) \to \chi \) under \( H_0(C) \) as \( N \to \infty \).

To show that \( \chi \) has unit expectation, we note that the trace in the denominator of \( F_N(C) \) accounts for the asymptotic expectation of the quadratic form term: the expectation of a quadratic form \( \text{tr}(X'AX) \) for a random vector \( X \) with mean \( \mu = E(X) \) and covariance matrix \( \Sigma = \text{cov}(X) \) equals \( E(X'AX) = \text{tr}(A \Sigma) + \mu' A \mu \); see Corollary 3.2b.1.60 In our situation, where \( A = T \) and \( T p = 0 \) under \( H_0(C) \), the asymptotic expectation of \( N \hat{p}' T \hat{p} \) under \( H_0 \) thus simplifies to \( tr(TV) \).

On the other hand, under the alternative hypothesis \( H_0(C) \), we have that

\[
F_N(T) = \frac{N}{tr(T \hat{V}_N)} \hat{p}' T \hat{p} = \frac{N}{tr(T \hat{V}_N)} (\hat{p} - p)' T (\hat{p} - p) + \frac{N}{tr(T \hat{V}_N)} (2 \hat{p} - p)' T p
\]

The first term on the right-hand side converges in distribution to \( \chi \) as \( N \to \infty \). Of the second term, \( (2 \hat{p} - p)' T p \) converges to the positive number \( p' T p \). Thus, an application of Slutsky’s lemma yields that \( F_N(T) \to \infty \) as \( \min n \to \infty \).

1.5 Proof of Theorem 3

This conditional central limit theorem holds because for each wild bootstrapped Kaplan–Meier estimator, a conditional central limit theorem holds separately. In particular, an appropriate wild bootstrap resampling technique for independently right-censored competing risks and survival data has been suggested in the presence of ties.45 They are combined with the help of the continuous mapping theorem and also with the consistency of the wild bootstrapped covariance estimators which were shown in that same paper. Hence, the continuous mapping theorem implies that the conditional distribution of \( F_N(T) \) given the data converges weakly in probability to the distribution of \( \chi \).
2 Asymptotic covariance matrix $V$ of $\sqrt{N}(\hat{p} - p)$

The covariance matrix $V$ in Theorem 1 has the compact form

$$V = \frac{1}{d^2} \sum_{j=1}^{d} \frac{1}{\lambda_j} \int \int \Gamma_j^{\pm} d\tilde{S} dS' + \int \int \text{diag} \left( \frac{1}{\lambda_i} \Gamma_i^{\pm} \right)_{i=1}^{d} d\tilde{S} dS'$$

$$- \frac{1}{d} \int \int \left( \frac{1}{\lambda_i} \Gamma_i^{\pm} \right)_{i=1}^{d} d\tilde{S} dS' - \left( \frac{1}{d} \int \int \left( \frac{1}{\lambda_i} \Gamma_i^{\pm} \right)_{i=1}^{d} d\tilde{S} dS' \right)'$$

$$= \frac{1}{d^2} \int \int \Gamma d\text{diag}(\lambda)^{-1} \Gamma d\tilde{S} dS' + \text{diag}(\lambda)^{-1} \int \int \text{diag}(\Gamma^{\pm}) d\tilde{S} dS'$$

$$- \frac{1}{d} \int \int \text{diag}(\lambda)^{-1} \Gamma d\tilde{S} dS' - \left( \frac{1}{d} \int \int \text{diag}(\lambda)^{-1} \Gamma d\tilde{S} dS' \right)'$$

where $\Gamma = (\Gamma_i)_{i=1,...,d}$ is aggregates the asymptotic covariance functions of all Kaplan–Meier estimators.