Received: 23 September 2020 | Accepted: 16 September 2021

DOI: 10.1111/biom.13575

BIOMETRIC METHODOLOGY

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

CASANOVA: Permutation inference in factorial survival designs

Marc Ditzhaus[1] | Jon Genuneit[2] | Arnold Janssen[3] | Markus Pauly[1]

[1] Department of Statistics, TU Dortmund University, Dortmund, Germany

[2] Pediatric Epidemiology, Department of Pediatrics, Leipzig University, Leipzig, Germany

[3] Mathematical Institute, Heinrich-Heine University Duesseldorf, Duesseldorf, Germany

Correspondence
Marc Ditzhaus, Department of Statistics, TU Dortmund University, Dortmund, Germany.
Email: marc.ditzhaus@tu-dortmund.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Numbers: 271512359, PA-2409 5-1

Abstract

We propose inference procedures for general factorial designs with time-to-event endpoints. Similar to additive Aalen models, null hypotheses are formulated in terms of cumulative hazards. Deviations are measured in terms of quadratic forms in Nelson–Aalen-type integrals. Different from existing approaches, this allows to work without restrictive model assumptions as proportional hazards. In particular, crossing survival or hazard curves can be detected without a significant loss of power. For a distribution-free application of the method, a permutation strategy is suggested. The resulting procedures' asymptotic validity is proven and small sample performances are analyzed in extensive simulations. The analysis of a data set on asthma illustrates the applicability.

KEYWORDS

additive Aalen model, factorial designs, local alternatives, oncology, right censoring

1 | INTRODUCTION

Kristiansen (2012) reviewed 175 studies with time-to-event endpoints published in five renowned journals. In 47% of these studies, crossing survival curves were present. The alarming observation of his review was: "Among studies with survival curve crossings, Cox regression was performed in 66% and log-rank-test in 70% of the studies." Under the assumption of proportional hazards, the log-rank test and the Cox regression are indeed very powerful tools. Otherwise, however, log-rank tests significantly lose power, and Cox regressions cannot be interpreted appropriately, in particular, when the survival curves cross. Thus, as stated by Bouliotis and Billingham (2011): "There is a need in the clinical community to clarify methods that are appropriate when survival curves cross." This especially holds in oncology, where crossing survival curves are frequently observed (Gahrton et al., 2013; Smith et al., 2014) due to a delayed treatment effect of immunotherapy (Mick and Chen, 2015).

In the two-sample setting, various inferences methods have been designed to detect nonproportional hazard alternatives, especially crossing curves. We refer to Li et al. (2015) for a review up until the year 2014, where a two-stage procedure (Qiu and Sheng, 2008) showed the most convincing performance. Recently, Fernández and Rivera (2020), Gorfine et al. (2020), Liu et al. (2020) proposed methods based on reproducing kernel Hilbert spaces, sample-space partitioning, and the area under the curve. Other tempting approaches are extensions of weighted log-rank tests (Gill, 1980; Andersen et al., 1993; Bathke et al., 2009), for which the power is optimized for certain nonproportional hazard alternatives by adding a respective weight function. For example, the weights of Harrington and Fleming (1982) can be used for late treatment effects, as recently recalled by Su and Zhu (2018). Such weights probably would have helped Jacobs et al.

(2016) to confirm their initial assumption in an ovarian cancer screening trial. Instead, they used the log-rank test and stated: "The main limitation of this trial was our failure to anticipate the late effect of screening in our statistical design. Had we done so, the weighted log-rank test could have been planned in line with many other large cancer screening trials." This quote illustrates the problem of the (weighted) log-rank tests: they are designed for specific alternatives and prior knowledge is needed to choose the optimal weight. To overcome this selection problem, Brendel et al. (2014) introduced a combination approach of several weights leading to procedures with broader power functions. Recently, their approach was revisited and simplified leading to computationally more efficient test versions (Ditzhaus and Pauly, 2019; Ditzhaus and Friedrich, 2020).

As "evaluating more than 1 new intervention concurrently increases the chances of finding an effective intervention" (Juszczak et al., 2019), we aim to extend their idea to more general factorial designs. This way we not only address the multiarm problem with crossing survival curves for which only a handful of relevant methods exist (Bathke et al., 2009; Chen et al., 2016; Liu and Yin, 2017; Gorfine et al., 2020) but also develop methods that fully exploit the structure of factorial survival designs. To the best of our knowledge, the proposed methods are the first to tackle the problem of nonproportional hazards in general factorial designs. Such methods will not only allow the detection of main factor effects but also infer potential interaction effects (Kurz et al., 2015) as, for example, also stated by Lubsen and Pocock (1994): "it is desirable for reports of factorial trials to include estimates of the interaction between the treatments." In contrast to Cox, Aalen, or Cox-Aalen regression models (Cox, 1972; Scheike and Zhang, 2003), there is no need to introduce multiple dummy variable for nominal factors (e.g., treatments) in the factorial design setup, which is even favorable in uncensored situations (Green, 2012). In the context of survival data, just a few nonparametric methods account for factorial designs: the approaches of Akritas and Brunner (1997), which require a strong assumption on the underlying censoring distribution that is often too strict from a practical point of view, and the procedures of Dobler and Pauly (2020) and Ditzhaus et al. (2021) formulating null hypotheses in terms of certain concordance effects, that restricts their analysis to a pre-specified time range $[0, \tau]$, and medians, respectively. All three approaches are not flexibly adaptable to detect certain crossing structures.

In contrast, we follow the spirit of the additive model of Aalen (1980) and formulate our null hypotheses utilizing cumulative hazard functions. Inspired by Neuhaus (1993) and Janssen and Mayer (2001), we derive critical values employing a permutation procedure. Under exchangeable data, for example, when the survival and the censoring distributions are the same over all groups, respectively, the permutation strategy leads to a finitely exact test under the null hypothesis (Lehmann and Romano, 2006). It is, however, less known that valid permutation tests can also be constructed without this restrictive assumption. In fact, to overcome the problem of potentially different censoring distributions, Wang et al. (2010) suggested a permutation imputation strategy. In detail, they proposed to impute censored observations from the conditional Kaplan–Meier-estimates for the groups' censoring distributions whenever they change the group during the permutation step. Recently, Gorfine et al. (2020) followed this strategy to derive a k-sample procedure. But the permutation imputation method does not tackle the problem of potential different survival curves, which may occur in factorial designs, nor preserves the favorable property of permutation tests being finitely exact under exchangeability. To tackle both limitations, we follow a different permutation scheme, namely, the idea of studentized permutation tests. These were mainly explored for testing means and other functionals in the two-sample case (Janssen and Pauls, 2003; Ditzhaus et al., 2021). Later, the concept of studentization was extended to one-way layouts by Chung and Romano (2013), and finally, reached its full potential under general factorial designs (Pauly et al., 2015; Ditzhaus et al., 2021). Thereof, only Ditzhaus et al. (2021) treated factorial survival models, but with a less flexible median-based approach. Therefore, the aims of the present paper are to derive a permutation procedure:

(a) without any restrictive assumption on the censoring distribution or the time range,
(b) with reasonable power under proportional hazards and crossing curve scenarios,
(c) for general factorial designs allowing the study of main and interaction effects, and
(d) being asymptotically valid with satisfactory small sample size performance.

This will be achieved by combining a weighted combination approach with the studentized permutation strategy.

In Section 2, we introduce the survival model and the null hypotheses formulated in terms of cumulative hazard rate functions. To test for certain main or interaction effects, we propose respective Wald-type statistics based on weighted Nelson–Aalen-type integrals, see Section 3. We prove their asymptotic validity and derive their power behavior under local alternatives. Motivated by the latter, we suggest to combine different weight functions into a joint Wald-type statistic to obtain a powerful method for various alternatives simultaneously, for example, proportional hazards and crossing survival curves. Respective permutation versions of them, promising a better finite sample performance, are shown to be asymptotically exact in Section 4. A simulation study presented in Section 5 reveals an actual improvement when using the permutation approach and show that the combination strategy actually results in a powerful test for alternatives with proportional hazards and with crossing curves. Finally, the tests' applicability is illustrated by analyzing data from a recent study on asthma.

2 | THE SETUP

Our general survival model is given by mutually independent positive random variables

$$T_{ji} \sim F_j, \quad C_{ji} \sim G_j \quad (j = 1, \dots, k;\ i = 1, \dots, n_j), \tag{1}$$

where T_{ji} is the actual survival time with continuous distribution function F_j and C_{ji} denotes the corresponding right-censoring time with continuous distribution function G_j. This setup allows the consideration of simple one-way but also of higher way layouts. For illustration, consider a two-way design with factors B (having b levels) and C (possessing c levels). In this scenario, we set $k = b \cdot c$ and split up the group index j into $j = (j_B, j_C)$ for $j_B = 1, \dots, b$ and $j_C = 1, \dots, c$. More complex designs, for example, hierarchical designs with nested factors, can be incorporated into this framework as well, see Dobler and Pauly (2020) for more details.

Based on the observation time $X_{ji} = \min(T_{ji}, C_{ji})$ and its censoring status $\delta_{ji} = I\{X_{ji} = T_{ji}\}$, where $I(\cdot)$ denotes the indicator function, we like to infer hypotheses formulated in terms of the cumulative hazard rate functions $A_j(t) = \int_0^t (1 - F_j)^{-1} F_j \ (t \geq 0)$:

$$\mathcal{H}_0 : \boldsymbol{HA} = \boldsymbol{0}_d, \quad \boldsymbol{A} = (A_1, \dots, A_k)', \tag{2}$$

where $\boldsymbol{H} \in \mathbb{R}^{d \times k}$ is a contrast matrix, that is, $\boldsymbol{H1_k} = \boldsymbol{0}_d$, and $\boldsymbol{0}_d$ as well as $\boldsymbol{1}_d$ are vectors in $\mathbb{R}^d$ consisting of 0's and 1's only. Here and subsequently, we use the following standard matrix notation: $\boldsymbol{B}'$ is the transpose and $\boldsymbol{B}^+$ is the Moore–Penrose inverse of a matrix $\boldsymbol{B}$. The contrast matrix in (2) is chosen in regard to the underlying question of interest. For example, in a one-way layout, the null hypothesis $\mathcal{H}_0 : \{A_1 = \cdots = A_k\} = \{\boldsymbol{P}_k \boldsymbol{A} = \boldsymbol{0}_k\}$ of no group effect can be expressed in terms of the contrast matrix $\boldsymbol{P}_k = \boldsymbol{I}_k - (\boldsymbol{J}_k / k)$, where $\boldsymbol{I}_k$ is the $k \times k$-dimensional unity matrix and $\boldsymbol{J}_k = \boldsymbol{1}_k' \boldsymbol{1}_k \in \mathbb{R}^{k \times k}$ consists of 1 only.

Switching to a two-way layout ($k = b \cdot c$) with the factors B (having b levels) and C (with c levels), the relevant matrices are $\boldsymbol{H}_B = \boldsymbol{P}_b \otimes (\boldsymbol{J}_c / c)$, $\boldsymbol{H}_C = (\boldsymbol{J}_b / b) \otimes \boldsymbol{P}_c$ and $\boldsymbol{H}_{BC} = \boldsymbol{P}_b \otimes \boldsymbol{P}_c$, where $\otimes$ is the Kronecker product. They can be used to check the null hypotheses

- *No main effect B:* $\{\boldsymbol{H}_B \boldsymbol{A} = \boldsymbol{0}_k\} = \{\bar{A}_{1\cdot} = \cdots = \bar{A}_{b\cdot}\}$.
- *No main effect C:* $\{\boldsymbol{H}_C \boldsymbol{A} = \boldsymbol{0}_k\} = \{\bar{A}_{\cdot 1} = \cdots = \bar{A}_{\cdot c}\}$.
- *No interaction effect:* $\{\boldsymbol{H}_{BC} \boldsymbol{A} = \boldsymbol{0}_k\} = \{\bar{A}_{\cdot\cdot} - \bar{A}_{\cdot j_C} - \bar{A}_{j_B \cdot} + A_{j_B j_C} = 0\}$.

Here, $\bar{A}_{j_B \cdot}$, $\bar{A}_{\cdot j_C}$, and $\bar{A}_{\cdot\cdot}$ are the means over the dotted indices. Suppose for a moment that hazard rates $\alpha_j(t) = dA_j(t)/dt \quad (t \geq 0)$ exist. Having the additive Aalen model in mind, we can rewrite the null hypotheses more lucidly by decomposing the hazard rate $\alpha_j = \alpha_{j_B j_C}$ into $\alpha_{j_B j_C}(t) = \alpha_0(t) + \beta_{j_B}(t) + \gamma_{j_C}(t) + (\beta\gamma)_{j_B j_C}(t)$ with side conditions $\sum_{j_B} \beta_{j_B} = \sum_{j_C} \gamma_{j_C} = \sum_{j_B} (\beta\gamma)_{j_B j_C} = \sum_{j_C} (\beta\gamma)_{j_B j_C} = 0$. Then we can rewrite $\{\boldsymbol{H}_C \boldsymbol{A} = \boldsymbol{0}_k\} = \{\gamma_{j_C} = 0 \text{ for all } j_C\}$ or $\{\boldsymbol{H}_{BC} \boldsymbol{A} = \boldsymbol{0}_k\} = \{(\beta\gamma)_{j_B j_C} = 0 \text{ for all } j_B, j_C\}$. More complex designs, for example, crossed three- and higher-way layouts, can be treated similarly, see the Supporting Information for details.

As in analysis-of-variance settings (Pauly et al., 2015), it is preferable to work with the projection matrix $\boldsymbol{T} = \boldsymbol{H}'(\boldsymbol{HH}')^+ \boldsymbol{H}$ over $\boldsymbol{H}$ itself. Beside from being unique, $\boldsymbol{T}$ is symmetric and idempotent, and describes the same null hypothesis as $\boldsymbol{H}$ does. We will therefore work with $\boldsymbol{T}$ when formulating our testing procedure. In addition, we need the usual counting process notation. Thus, let $N_j(t) = \sum_{i=1}^{n_j} I\{X_{ji} \leq t, \delta_{ji} = 1\}$ be the number of observed events within group j until time t and introduce $Y_j(t) = \sum_{i=1}^{n_j} I\{X_{ji} \geq t\}$, the number of individuals being at risk just before t in the same group. These processes allow us to define the Nelson–Aalen estimator for A_j given by $\widehat{A}_j(t) = \int_0^t I\{Y_j(s) > 0\} / Y_j(s) \, \mathrm{d}N_j(s) \ (j = 1, \dots, k;\ t \geq 0)$. For our purposes, the Kaplan–Meier estimator is only required for all $n = \sum_{j=1}^k n_j$ observations without distinguishing between the groups. This pooled version $\widehat{F}$ can be expressed in terms of the pooled counting processes $N = \sum_{j=1}^k N_j$ and $Y = \sum_{j=1}^k Y_j$ by $1 - \widehat{F}(t) = \prod_{(j,i): X_{ji} \leq t} [1 - \{\Delta N(X_{ji}) / Y(X_{ji})\}] (t \geq 0)$, where $\Delta N(t) = N(t) - N(t-)$ denotes the increment at t. In the same way, $\widehat{A}(t) = \int_0^t I\{Y > 0\} / Y \, \mathrm{d}N \ (t \geq 0)$ denotes the pooled Nelson–Aalen estimator.

3 | ASYMPTOTIC RESULTS

3.1 | Wald-type test

Throughout, we assume nonvanishing groups $n_j / n \to \kappa_j \in (0, 1)$ as $\min(n_j : j = 1, \dots, k) \to 0$. Moreover, we exclude the trivial case of purely censored observations in any of the groups by assuming that $F_j(t) > 0$ and $G_j(t) < 1$ for all $j = 1, \dots, k$ and some $t > 0$.

Weighted log-rank statistics (Andersen et al., 1993, e.g.) of the form

$$\left(\frac{n}{n_1 n_2}\right)^{1/2} \int_0^\infty \widetilde{w}\{\widehat{F}_n(t-)\}\frac{Y_1(t)Y_2(t)}{Y(t)}\left\{\mathrm{d}\widehat{A}_1(t) - \mathrm{d}\widehat{A}_2(t)\right\},$$

will later build the fundament of our new test statistics. Here, $t \mapsto \widehat{F}_n(t-)$ is the left continuous version of $\widehat{F}_n$ and $\widetilde{w}$ is a weight function taken from the space $\mathcal{W}$ consisting of all continuous functions $\widetilde{w} : [0,1] \to \mathbb{R}$ of bounded variations with $\widetilde{w}(t) \neq 0$ for some $t \in [0,1]$. Fleming and Harrington (1991) considered a subclass of these weights having the shape $\widetilde{w}(t) = t^r(1-t)^g$ $(r, g \in \mathbb{N}_0)$, see their Definition 7.2.1. Setting $r = g = 0$, we obtain the log-rank test and the choice $(r, g) = (1, 0)$ leads to the Prentice–Wilcoxon test. In general, these weights can be used to prioritize (mid-)late, (mid-)early, or central times by choosing r, g appropriately. For our purposes, weights, for example, $\widetilde{w}(t) = 1 - 2t$, intersecting the x-axis are of special interest because they are designed for crossing hazard alternatives.

Having all these weights at hand, the question arises: which weight should be chosen? We address this question in detail in the next two sections, but first, we introduce the relevant components of the new test statistic. For ease of presentation, we restrict ourselves to nontrivial polynomial weights $\widetilde{w} \in \mathcal{W}$ covering the main relevant cases. However, more general weight functions can be treated analogously as discussed in the Supporting Information. First, we extend the weighted log-rank integrand to the present situation of multiple subgroups

$$w_n(t) = \widetilde{w}\{\widehat{F}_n(t-)\}\frac{Y_1(t)\ldots Y_k(t)}{nY(t)^{k-1}} \quad (t \geq 0), \tag{3}$$

and then define the Nelson–Aalen-type integral over $Z_{nj}(w_n) = n^{1/2}\int_0^\infty w_n(t)\,\mathrm{d}\widehat{A}_j(t)$. By standard martingale theory (Gill, 1980; Andersen et al., 1993), $\widetilde{Z}_{nj}(w_n) = n^{1/2}\int_0^\infty w_n(t)\{\mathrm{d}\widehat{A}_j(t) - \mathrm{d}A_j(t)\}$ is asymptotically $N\{0, \sigma_j^2(w)\}$-distributed with $\sigma_j^2(w) > 0$. The latter can be consistently estimated by $\widehat{\sigma}_j^2(w_n) = n\int_0^\infty \{w_n(t)^2/Y_j(t)\}\,\mathrm{d}\widehat{A}_j(t)$, see the Supporting Information. It follows (Rao and Mitra, 1971, Th. 9.2.2) that, under $\mathcal{H}_0 : \boldsymbol{TA} = \boldsymbol{0}_k$, the Wald-type statistic

$$S_n(w_n) = [\boldsymbol{TZ}(w_n)]'(\boldsymbol{T}\widehat{\boldsymbol{\Sigma}}(w_n)\boldsymbol{T})^{+}\boldsymbol{TZ}(w_n),$$
$$\widehat{\boldsymbol{\Sigma}}(w_n) = \mathrm{diag}\{\widehat{\sigma}_1^2(w_n), \ldots, \widehat{\sigma}_k^2(w_n)\} \tag{4}$$

is asymptotically χ_f^2-distributed with $f = \mathrm{rank}(\boldsymbol{T})$ degrees of freedom. To motivate a sensible combination of weights, we study the asymptotic power of $S_n(w_n)$ under local alternatives.

3.2 | Local alternatives

To this end, we start with a fixed null setting $\boldsymbol{A} = (A_1, \ldots, A_k)^T$ with $\boldsymbol{TA} = \boldsymbol{0}_k$ and corresponding hazard rates $\alpha_j(t) = \mathrm{d}A_j(t)/\,\mathrm{d}t$ $(t \geq 0)$. Disturbing them as follows, we get a local alternative $(A_{1n}, \ldots, A_{nk})$ tending with a rate of $n^{-1/2}$ to the null setting $\boldsymbol{A}$:

$$\frac{\alpha_{nj}(t)}{\alpha_j(t)} = 1 + n^{-1/2}\gamma_j(t) \quad (j = 1, \ldots, k; t \geq 0), \tag{5}$$

where the right hand side is nonnegative and $\int_0^t \gamma_j(u)\alpha_j(u)\,\mathrm{d}u \in \mathbb{R}$ for all $t \geq 0$ fulfilling $F_j(t) < 1$. To simplify the situation, we may restrict to perturbations

$$\gamma_j(t) = \theta_j\gamma\{F_0(t)\}. \tag{6}$$

in the same direction γ but with possibly different strengths $\theta_j > 0$. Here, F_0 is the limit function of the pooled Kaplan–Meier estimator, see the Supporting Information for its concrete shape. Moreover, we introduce $y_j = \kappa_j(1 - G_j)(1 - F_j)$ $(j = 1, \ldots, k)$, which is the limit of $n^{-1}Y_j$.

Theorem 1. *Under Assumption (5), $S_n(w_n)$ converges to a noncentral $\chi_f^2(\delta)$-distribution with $f = rank(\boldsymbol{T})$ and $\delta = (\boldsymbol{T\mu})'(\boldsymbol{T\Sigma T})^{+}\boldsymbol{T\mu}$, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)'$ and*

$$\mu_j = \int_0^\infty \widetilde{w}\{F_0(t)\}\frac{y_1(t)\ldots y_k(t)}{y(t)^{k-1}}\gamma_j(t)\,\mathrm{d}A_j(t), \quad y = \sum_{j=1}^k y_j.$$

The effect of the weight function on the power under certain local alternatives can be illustrated best for the k-sample setting under (6). In this case,

$$\delta = \left[\int_0^\infty \widetilde{w}\{F_0(t)\}\gamma\{F_0(t)\}\frac{y_1(t)\ldots y_k(t)}{y(t)^{k-1}}\,\mathrm{d}A_1(t)\right]^2 \times (\boldsymbol{T\theta})'(\boldsymbol{T\Sigma T})^{+}\boldsymbol{T\theta},$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$ and $\boldsymbol{\Sigma} = \mathrm{diag}\{\sigma_1^2(\widetilde{w}), \ldots, \sigma_k^2(\widetilde{w})\}$. Consequently, choosing $\widetilde{w}$ as a multitude of γ leads to the highest value for δ and, consequently, to the highest power of $S_n(w)$. However, the direction γ of the departure from the null hypothesis is unknown and again the question arises: how to choose $\widetilde{w}$? The task of finding the optimal $\widetilde{w}$ is impossible. The most popular choice is the log-rank test ($\widetilde{w} \equiv 1$) that, however, lacks to detect crossing hazard departures. To compensate for that, we follow Ditzhaus and Friedrich (2020) and suggest to combine the log-rank weight and a weight for crossing hazard alternatives, for example, $\widetilde{w}(x) = 1 - 2x$ $(0 \leq x \leq 1)$, into a joint Wald-type statistic. In general, the new approach is not restricted to these two weights and even more than two weights can be combined that is particularly useful if there is prior knowledge of alternatives to be discovered.

3.3 | Combination of different weights

Let us start with an arbitrary number of prechosen weights $\widetilde{w}_1, \ldots, \widetilde{w}_m$ corresponding to alternatives of interest, for example, proportional, late, early, or crossing hazards. Moreover, let $w_{n1}, \ldots, w_{nk}$ be the corresponding integrands of the form (3) for the Nelson–Aalen-type integrals. To exclude redundant cases, as too similar or even equal weights, we follow Ditzhaus and Pauly (2019) and Ditzhaus and Friedrich (2020) and restrict to weights fulfilling the following.

Assumption 1. *Let $\widetilde{w}_1, \ldots, \widetilde{w}_m \in \mathcal{W}$ be linearly independent, nontrivial polynomials.*

The basic idea is to combine $\boldsymbol{Z}_n(w_{n1}), \ldots, \boldsymbol{Z}_n(w_{nm})$ into one joint Wald-type statistic. For this purpose, we introduce the block diagonal matrix $\boldsymbol{T}^{(m)} = \mathrm{diag}(\boldsymbol{T}, \ldots, \boldsymbol{T}) \in \mathbb{R}^{km \times km}$. As $Z_{nj}(w_{nr})$ and $Z_{nj}(w_{nr'})$ are highly dependent, the vectors $\boldsymbol{Z}_n(w_{nr})$ and $\boldsymbol{Z}_n(w_{nr'})$ are so as well. Thus, the covariance matrix estimator required for the joint Wald-type statistic is not a simple diagonal matrix as in (4). In fact, the updated estimator has a block matrix representation $\widehat{\boldsymbol{\Sigma}} = (\widehat{\boldsymbol{\Sigma}}^{(rr')})_{r,r'=1,\ldots,m}$, where each submatrix $\widehat{\boldsymbol{\Sigma}}^{(rr')} = \mathrm{diag}(\widehat{\sigma}_1^{2,(rr')}, \ldots, \widehat{\sigma}_k^{2,(rr')})$ is a diagonal matrix with entires $\widehat{\sigma}_j^{2,(rr')} = n \int_0^\infty \{w_{nr}(t) w_{nr'}(t)/Y_j(t)\}\, \mathrm{d}\widehat{A}_j(t)$. To sum up, we obtain the following updated Wald-type statistic:

$$S_n = (\boldsymbol{T}^{(m)}\boldsymbol{Z}_n)'(\boldsymbol{T}^{(m)}\widehat{\boldsymbol{\Sigma}}\boldsymbol{T}^{(m)})^{+}\boldsymbol{T}^{(m)}\boldsymbol{Z}_n,$$

$$\boldsymbol{Z}_n = \{\boldsymbol{Z}_n(w_{n1})', \ldots, \boldsymbol{Z}_n(w_{nk})'\}'. \tag{7}$$

Theorem 2. *Under Assumption 1 and $\mathcal{H}_0 : \boldsymbol{TA} = \boldsymbol{0}_k$, S_n tends to a χ^2_f-distribution with $f = m \cdot rank(T)$ degrees of freedom as $n \to \infty$.*

By Theorem 2, an asymptotically exact test $\phi_n = I\{S_n > \chi^2_{f,\alpha}\}$, that is, $E_{\mathcal{H}_0}(\phi_n) \to \alpha$ as $n \to \infty$, is derived by comparing the joint Wald-type statistic S_n with the $(1-\alpha)$-quantile $\chi^2_{f,\alpha}$ of the χ^2_f-distribution. However, simulation results from Section 5 reveal a very conservative behavior of ϕ_n under small sample sizes. To tackle this problem, we suggest a permutation strategy leading to a better finite sample performance as can be seen in Section 5.

4 | PERMUTATION TEST

Resampling methods and, in particular, permutation procedures are well-accepted tools to improve the finite sample performance of asymptotic tests. The advantage of permuting over other resampling methods is the finite sample exactness of the test under exchangeable data, that is, under the restrictive null hypothesis $\widetilde{\mathcal{H}}_0 : A_1 = \cdots = A_k, G_1 = \cdots = G_k$ in our scenario. At the same time, the asymptotic exactness of the test beyond exchangeability can often be transferred to its permutation counterpart when working with studentized statistics, as the present joint Wald-type statistic. That is why we promote the following permutation strategy for our setting: To obtain a permutation sample $\{(X^\pi_{ji}, \delta^\pi_{ji}) : j = 1, \ldots, k; i = 1, \ldots, n_j\}$, we randomly interchange the group memberships of the observation pairs (X_{ji}, δ_{ji}). With this, we calculate the permutation version of the joint Wald-type statistic $S^\pi_n = S_n((X^\pi_{ji}, \delta^\pi_{ji})_{j,i})$.

Theorem 3. *Under Assumption 1, the permutation counterpart S^π_n of S_n always asymptotically mimics its null distribution, that is, under $\mathcal{H}_0$ as well as under fixed and local alternatives (5), we have $\sup_{t \in [0,\infty)} |P\{S^\pi_n \leq t \mid (X_{ji}, \delta_{ji})_{ji}\} - \chi^2_{m \cdot rank(\boldsymbol{T})}(-\infty, t]| \xrightarrow{p} 0$ as $n \to \infty$.*

As, for example,, elaborated by Dobler and Pauly (2014), care has to be taken when applying permutation techniques in combination with continuous martingale theory. To prove Theorem 3, we therefore transfer the proof technique of Ditzhaus and Janssen (2020) relying on discrete martingales to the present setup.

Theorem 3 allows to compare the joint Wald-type statistic S_n with the $(1-\alpha)$-quantile $c^\pi_{n,\alpha}$ of $t \mapsto P\{S^\pi_n \leq t \mid (X_{ji}, \delta_{ji})_{ji}\}$ (instead of the asymptotic χ^2_f-quantile). This results in the permutation test $\phi^\pi_n = I\{S_n > c^\pi_{n,\alpha}\}$ having type-I error as well as power behavior under fixed and local alternatives (Janssen and Pauls, 2003, Lemma 1).

5 | SIMULATIONS

5.1 | Simulation setup

In addition to the asymptotic findings, we conduct a simulation study to examine the tests' small sample performance.

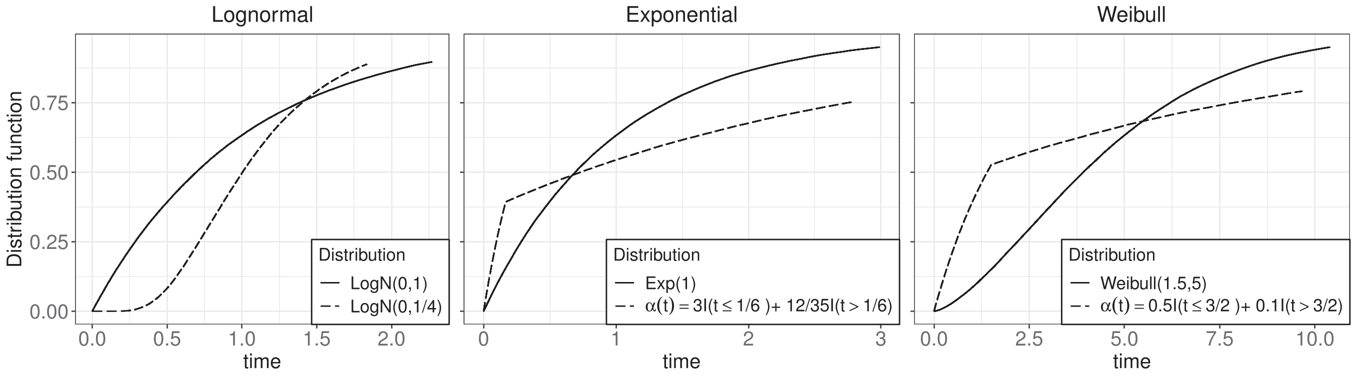

FIGURE 1 Crossing survival curve alternatives

5.1.1 | Designs

First, we consider a 2×2 layout with factors B and C and infer the null hypotheses $\{\boldsymbol{H}_C\boldsymbol{A} = \mathbf{0}_4\}$ of no main effect of C and of no interaction effect $\{\boldsymbol{H}_{BC}\boldsymbol{A} = \mathbf{0}_k\}$. Here, we discuss balanced and unbalanced scenarios given by sample size vectors $\mathbf{n}_{\text{bal}} = (n_{11}, n_{12}, n_{21}, n_{22}) = (15, 15, 15, 15)$ and $\mathbf{n}_{\text{unb}} = (10, 20, 10, 20)$ as well as their multiples $K\mathbf{n}_{\text{bal}}$ and $K\mathbf{n}_{\text{unb}}$ for $K = 2, 4$, respectively. We also study four different censoring settings $\mathbf{cr}_1 = (cr_{11}, cr_{12}, cr_{21}, cr_{22}) = (7\%, 12\%, 12\%, 7\%)$ (*low*), $\mathbf{cr}_2 = (29\%, 38\%, 25\%, 35\%)$ (*medium*), $\mathbf{cr}_3 = (12\%, 38\%, 7\%, 29\%)$ (*low/medium*), $\mathbf{cr}_4 = (55\%, 65\%, 55\%, 65\%)$ (*high*).

5.1.2 | Survival times

Under the null hypotheses, the survival times are simulated according to the same distribution in each (sub-)group, where we consider the standard Exponential distribution Exp(1), a Weibull distribution Weibull(1.5, 5) with parameters $(\lambda_{\text{shape}}, \lambda_{\text{scale}}) = (1.5, 5)$ and a standard log-normal distribution $\text{LogN}(0, 1)$. To obtain relevant alternatives, we disturb these null settings by choosing three different crossing curve alternatives, see Figure 1, and two proportional hazard alternatives: $\text{Weibull}(1.5, 5(2.5)^{-2/3})$ and Weibull(1.5, 5) as well as Exp(2.5) and Exp(1) for group 1 and 2, respectively. In both cases, the respective hazard ratio equals 2.5. In the present two-way layout, the observations of the first subgroup, $(j_B, j_C) = (1, 1)$, for testing of no interaction effect, and of the first and third subgroups $(j_B, j_C) = (1, 1), (2, 1)$ for testing of no main effect C are generated according to these alternative distributions, whereas the remaining observations follow the respective null distribution.

5.1.3 | Censoring times

The censoring times are simulated by uniform distributions $\text{Unif}[0, U_j]$. The upper limit U_j of the interval in group j is determined by a Monte-Carlo simulation such that the average censoring rate $P(T_{j1} > C_{j1}) = \int_0^\infty \min\{x/U_j, 1\}\,\mathrm{d}F_j(x)$ equals the prechosen rate cr_j. As suggested by a referee, we also consider the situation of heterogeneous censoring patterns to study the robustness of the permutation procedure under the null hypothesis. For that purpose, we consider two censoring distributions depending on the level j_C of the second factor C. For $j_C = 2$, we keep the above censoring mechanism, while for $j_C = 1$, we replace it by respective exponentially distributed censoring times, where again the appropriate parameter is determined by Monte-Carlo simulation.

5.1.4 | Methods

For the type-I error rate and power comparisons, we include the proposed singly-weighted Wald-type tests with the log-rank weight $\widetilde{w}_1(t) = 1$ and a crossing weight $\widetilde{w}_2(t) = 1 - 2t$, respectively, as well as the joint Wald-type test combining these two. We compare the performance of the respective asymptotic and permutation procedures. The simulation study is complemented by the well-known Aalen- and Cox-regression methods. For both, we include the main effects when testing for no main effects, and additionally, the interaction effect when testing for no interaction. All simulations are conducted by means of the computing environment R (R Core Team, 2020), version 3.6.2. For each setting, $N_{\text{sim}} = 5000$ simulation runs and $N_{\text{perm}} = 1999$ permutation iterations were generated.

TABLE 1 Type-I error rates in % (nominal level $\alpha = 5\%$) in the 2×2-layout under the mixed censoring setting

			Low cens.		Low/Medium		Medium cens.		High cens.	
Effect	**Distr**	n	**Asy**	**Per**	**Asy**	**Per**	**Asy**	**Per**	**Asy**	**Per**
Interaction	Exp	$\mathbf{n}_{unb}$	3.3	**4.8**	3.2	**5.0**	2.9	**4.8**	2.7	**4.6**
		$\mathbf{n}_{bal}$	3.7	**5.0**	3.9	**4.9**	4.2	**5.0**	3.6	**5.3**
		$2\mathbf{n}_{unb}$	**4.4**	**5.2**	3.7	**4.8**	4.3	**5.3**	3.6	**5.1**
		$2\mathbf{n}_{bal}$	**4.8**	**5.3**	4.0	**4.5**	**4.5**	**5.1**	4.2	**5.1**
		$4\mathbf{n}_{unb}$	**4.6**	**5.0**	**5.0**	**5.5**	**4.8**	**5.4**	4.3	**5.1**
		$4\mathbf{n}_{bal}$	**4.9**	**4.8**	**5.0**	**5.1**	**4.8**	**5.1**	**4.7**	**5.3**
	logN	$\mathbf{n}_{unb}$	3.4	**4.9**	3.7	**5.3**	3.6	**5.0**	2.7	**5.4**
		$\mathbf{n}_{bal}$	3.6	**4.7**	3.5	**4.7**	3.5	**4.5**	3.1	**4.6**
		$2\mathbf{n}_{unb}$	**4.5**	**5.2**	3.7	**4.7**	**4.5**	**5.4**	4.0	**5.1**
		$2\mathbf{n}_{bal}$	**4.8**	**5.2**	4.3	**4.9**	4.2	**4.6**	3.7	**4.9**
		$4\mathbf{n}_{unb}$	**4.6**	**4.8**	**4.6**	**5.3**	**4.9**	**5.3**	3.8	**4.4**
		$4\mathbf{n}_{bal}$	**4.7**	**4.8**	**4.7**	**5.0**	**4.6**	**4.8**	4.2	**4.7**
	Weibull	$\mathbf{n}_{unb}$	3.5	**4.9**	3.0	**5.0**	3.6	**5.4**	2.8	**5.3**
		$\mathbf{n}_{bal}$	3.6	**4.9**	3.1	**4.4**	3.5	**4.6**	3.0	**4.8**
		$2\mathbf{n}_{unb}$	4.3	**4.9**	4.0	**5.0**	4.3	**5.0**	3.7	**4.9**
		$2\mathbf{n}_{bal}$	**4.7**	**5.1**	4.0	**4.4**	4.2	**4.9**	4.3	**5.3**
		$4\mathbf{n}_{unb}$	**4.6**	**4.8**	**4.5**	**5.1**	**4.8**	**5.1**	**4.7**	**5.6**
		$4\mathbf{n}_{bal}$	**5.0**	**5.1**	**4.6**	**4.8**	**5.0**	**5.2**	4.2	**4.8**
Main	Exp	$\mathbf{n}_{unb}$	4.3	**4.9**	**4.7**	**5.3**	4.1	**4.7**	3.9	3.9
		$\mathbf{n}_{bal}$	3.5	**4.5**	3.5	**4.9**	4.0	**5.6**	3.7	**5.4**
		$2\mathbf{n}_{unb}$	**5.0**	**5.3**	**4.8**	**5.2**	4.3	**4.6**	**4.8**	**5.1**
		$2\mathbf{n}_{bal}$	**4.4**	**4.7**	**4.4**	**5.2**	**4.4**	**4.9**	**4.4**	**5.4**
		$4\mathbf{n}_{unb}$	**4.4**	**4.6**	**4.9**	**5.1**	**5.0**	**5.4**	**4.9**	**4.9**
		$4\mathbf{n}_{bal}$	**4.8**	**4.9**	**5.0**	**5.3**	**5.2**	**5.3**	**4.7**	**5.3**
	logN	$\mathbf{n}_{unb}$	4.2	**4.9**	**4.6**	**5.2**	3.7	**4.4**	4.3	**4.4**
		$\mathbf{n}_{bal}$	4.1	**5.1**	4.1	**5.5**	4.2	**5.5**	3.6	**5.4**
		$2\mathbf{n}_{unb}$	**5.0**	**5.4**	**5.4**	5.7	4.0	**4.4**	4.1	**4.5**
		$2\mathbf{n}_{bal}$	**4.4**	**4.7**	4.0	**4.6**	**5.0**	**5.4**	4.3	**5.6**
		$4\mathbf{n}_{unb}$	**4.9**	**5.3**	**4.6**	**4.9**	4.2	**4.5**	**4.7**	**4.9**
		$4\mathbf{n}_{bal}$	**5.0**	**5.1**	**4.4**	**4.6**	**4.9**	**5.1**	4.3	**4.9**
	Weibull	$\mathbf{n}_{unb}$	4.0	**4.8**	4.1	**4.7**	3.9	**4.6**	**4.7**	**4.9**
		$\mathbf{n}_{bal}$	4.1	**5.3**	3.3	**4.8**	3.6	**5.0**	3.3	**5.3**
		$2\mathbf{n}_{unb}$	**5.2**	**5.5**	**4.7**	**5.2**	4.2	**4.6**	**4.4**	**4.8**
		$2\mathbf{n}_{bal}$	4.0	**4.6**	**4.8**	**5.3**	**4.8**	**5.4**	**4.4**	**5.4**
		$4\mathbf{n}_{unb}$	**4.9**	**5.0**	**4.9**	**5.0**	**5.2**	**5.3**	**4.5**	**4.7**
		$4\mathbf{n}_{bal}$	**5.2**	**5.3**	**5.0**	**5.3**	**4.7**	**5.1**	**4.5**	**5.1**

Asy: joint Wald-type test based on the χ^2_f-approximation; Per: joint permutation Wald-type test.
Values inside the 95% binomial interval [4.4,5.6] are printed in bold.

5.2 | Simulation results

5.2.1 | Type-I errors

Table 1 presents the empirical sizes of the asymptotic joint Wald-type test and its permutation counterpart under the heterogeneous censoring setting, for example, uniform and exponential censoring. The results for the pure uniform censoring settings are presented in the Supporting Information and are slightly less extreme regarding the asymptotic test and equally convincing for the permutation approach. Besides, the Supporting Information contains all results for the Cox- and Aalen-regression, indicating a slight liberality for Cox and a slight conservativeness for

Aalen in case of the no interaction hypotheses and small sample sizes, and good control for both in the remaining settings. From Table 1, it is apparent that the asymptotic approach leads to rather conservative decisions for small sample sizes with values reaching down to 2.9%. In fact, out of the 48 small sample size settings, just in three of them the empirical size of the asymptotic test is within the 95% binomial confidence interval $[4.4\%, 5.6\%]$ for the estimated sizes. Doubling the sample sizes lead to an improvement of the asymptotic test's type-I error control. Now, just 6 out of 24 empirical sizes for the main effect but still 17 out of 24 empirical sizes for the interaction effect are outside the interval $[4.4\%, 5.6\%]$. For larger sample sizes $4\mathbf{n}_{\text{bal}}$ and $4\mathbf{n}_{\text{unb}}$, the empirical sizes all lie inside the interval for the low, low/medium, and medium censoring settings, and 7 out of 12 sizes lie inside the interval under high censoring. Overall, it can be seen that the conservativeness is more pronounced when the censoring is high. As the latter situation provides less information about the data, this observation is not surprising. In contrast, the permutation strategy leads to an accurate type-I error control even for small sample sizes: only two empirical size lie outside the confidence interval.

5.2.2 | Power

In Figure 2, all tests' power curves are displayed when testing for no interaction effects in the unbalanced sample size setting with sample size vectors $K\mathbf{n}_{\text{unb}}$, $K \in \{1, 2, 4\}$. The results for the balanced cases are comparable and, thus, shown in the Supporting Information. It is apparent that for the small sample size setting $\mathbf{n} = \mathbf{n}_{\text{unb}}$, the permutation joint Wald-type test leads to slightly higher power values than its respective asymptotic counterpart. The difference in terms of power becomes smaller or even vanishes for increased sample sizes. These observations are in line with the simulations under the null hypotheses, where the asymptotic joint Wald-type test exhibits a conservative behavior for smaller sample sizes. Comparing the joint Wald-type test with the two singly-weighted Wald-type tests and the two regression methods, we can draw two main conclusions: (1) Under the two proportional hazard alternatives, the Aalen- and Cox-regression as well as the singly-weighted Wald-type test with the log-rank weight leads to the highest power values. Their power curves are quite close, followed by the ones of the asymptotic and permutation joint Wald-type tests. The singly weighted Wald-type test with the crossing weight shows the lowest power values. (2) The prior findings completely change under the three crossing curve alternatives. Here, the asymptotic and permutation joint Wald-type tests exhibit the best power results. Although the singly-weighted Wald-type test with a crossing weight is slightly better under the Weibull distribution setting, the two joint Wald-type tests lead to significantly higher power values under the exponential and lognormal setups. In contrast to the proportional hazard alternatives, the two regression methods as well as the singly weighted Wald-type test with the log-rank weight lead to a rather poor detection rate. Under high censoring, the observations partially change. The first observation is that the detection rate of the regression methods and of the singly weighted Wald-type test with the log-rank weight improves significantly under the Weibull crossing setting, whereas it remains poor for the other two crossing settings. This can be explained by the rather late crossing in the Weibull scenario, see Figure 1, which is hardly observable under high censoring. Thus, this setting cannot be seen as a crossing situation anymore. The second observation is the higher power of the singly weighted Wald-type test with the crossing weight compared to the combination approach under the proportional hazard settings. Due to the high censoring, the pooled Kaplan–Meier estimator remains in average on a low level, that is, does not exceed 0.5 significantly. Thus, the crossing weight $\widetilde{w}_2(\widehat{F}(t)) = 1 - 2\widehat{F}(t)$ in (3) just emphasizes early times.

Summarizing the results, the combination approach offers a reasonable compromise with a convincing power performance under proportional hazards and crossing curves. It even outperforms the singly weighted Wald-type test with a crossing weight under two of the three crossing curve alternatives. This observation can be explained by interpreting the Wald-type statistic S_n as a projection, see Brendel et al. (2014) for details regarding the two-sample setting. The take-home message is that combining the weights $\widetilde{w}_1$ and $\widetilde{w}_2$ as well as combining $\widetilde{w}_1$ and $\widetilde{w}_3 = \widetilde{w}_2 + \lambda\widetilde{w}_1$ for some $\lambda \in \mathbb{R}$ results in the same statistic S_n. Although $\widetilde{w}_2$ is designed for crossings near to the center, the hazard rates in the exponential setting cross at a mid-early time, and thus, another crossing weight, for example, $\widetilde{w}_3 = \widetilde{w}_2 - 0.25$, would be more appropriate. But the choice between $\widetilde{w}_2$ or $\widetilde{w}_3$ does not affect the final result of the combination approach. To sum up, the advantages of the combination approach are that we neither need to choose between proportional and crossing hazard alternatives nor between different crossing points. For small sample sizes, we recommend the permutation test over its asymptotic counterpart due to the unstable type-I error rate behavior of the latter.

6 | ILLUSTRATIVE DATA ANALYSIS

We illustrate our approach using a longitudinal study on the well-established observations that (i) asthma occurs less often among children who grew up on farms

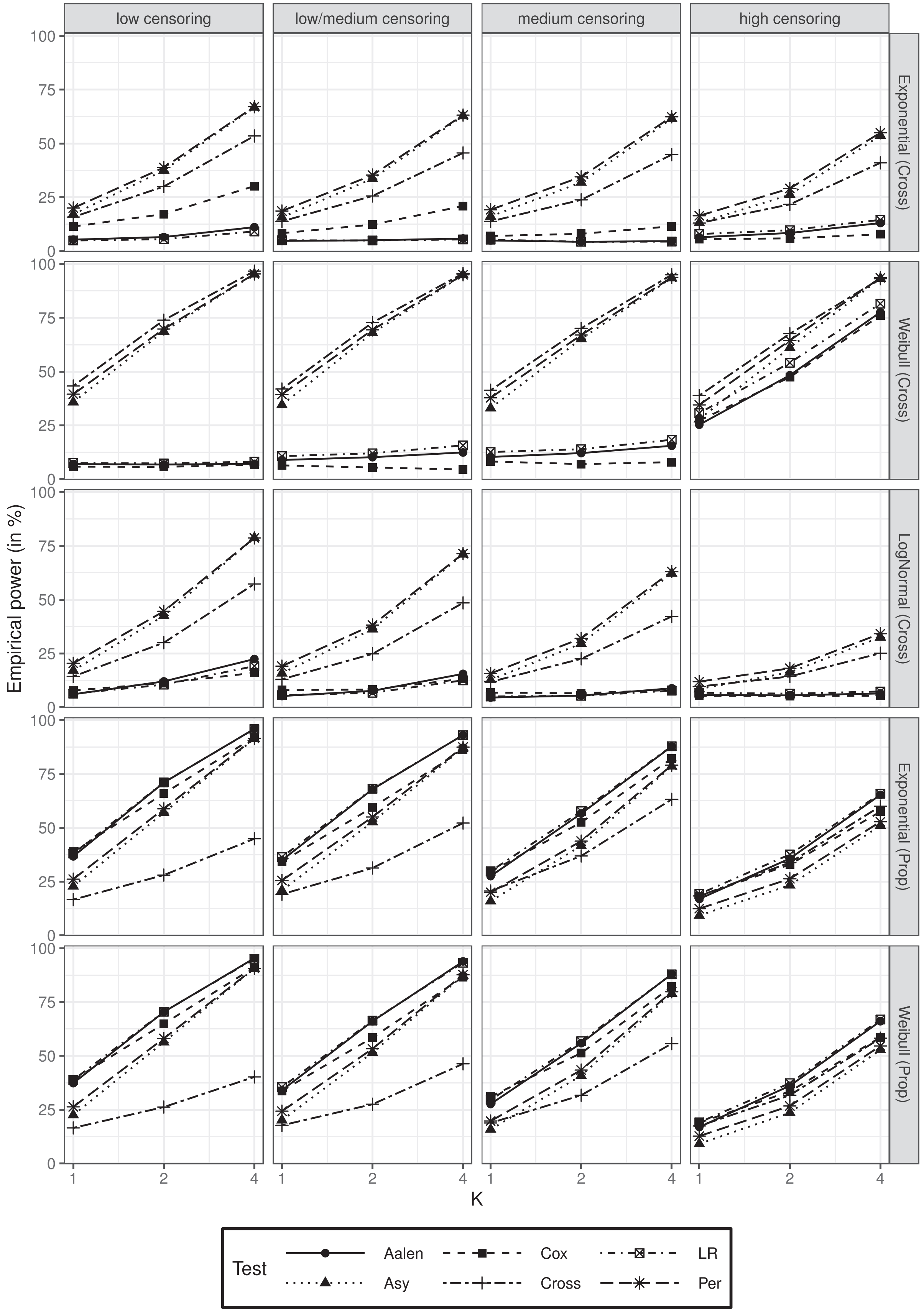

FIGURE 2 Power curves for testing of no interaction effect in the two-way layout using Aalen- and Cox-regression, the joint Wald-type test based on the χ_f^2-approximation (Asy), and the permutation approach (Per) as well as the singly weighted permutation tests based on w_1 (LR) and w_2 (Cross), respectively, under increasing unbalanced sample size $\boldsymbol{n} = K\boldsymbol{n}_{\text{unb}}$, $K = 1, 2, 4$

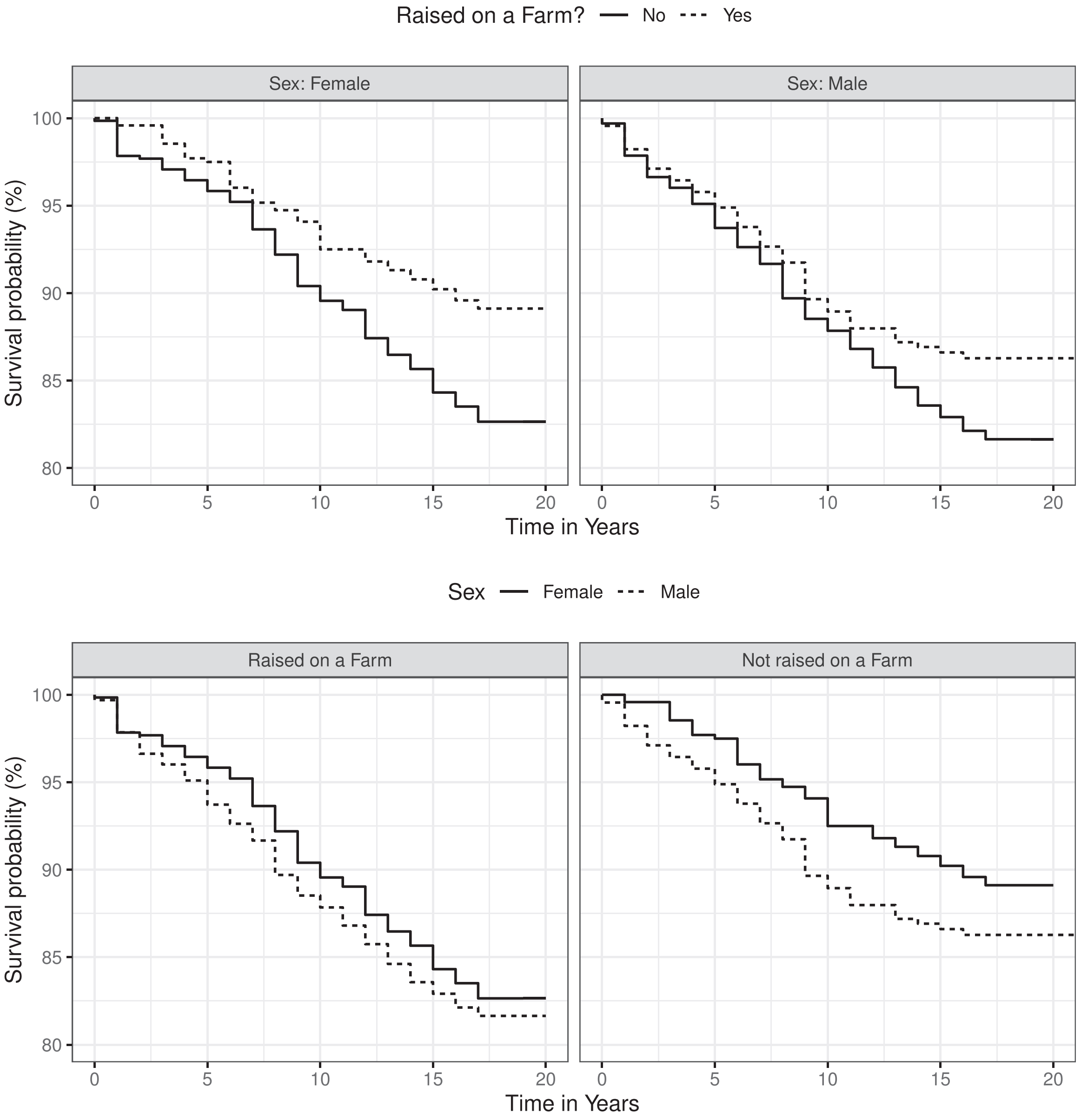

FIGURE 3 Kaplan–Meier curves for the asthma study

compared to children without contact to farming environments (Genuneit, 2012) and (ii) boys develop asthma more often in childhood compared to girls, whereas this ratio inverts during adolescence (Zein and Erzurum, 2015). Baseline assessment was a survey among elementary school children in southern Germany in the years 2006/07 (Genuneit et al., 2011). Interim analysis of the follow-up data until 2013 suggested a statistically significant interaction between child sex and living on a farm on the incidence of asthma (Genuneit, 2014). Here, we use all follow-up data until 2016, when the $n = 2230$ participants had reached 16–20 years of age. The data can be divided into four subgroups: 479 girls growing up on a farm (group 1, where 90.2% of the data are censored), 450 boys growing up on a farm (group 2, 86.7% censored), 648 girls not growing up on a farm (group 3, 84.9% censored), and 653 boys not growing up on a farm (group 4, 83.6% censoring). The Kaplan–Meier curves of these subgroups are displayed in Figure 3. The main questions of interest are: (1) Is there an effect of being raised on a farm? (2) Are there gender-specific differences? (3) Is there an interaction effect between these two factors? All three questions can be addressed using the presented combination approach by treating the data set as a 2×2 design and choosing the respective contrast matrices from Section 2. Due to rather large sample sizes, we apply the asymptotic Wald-type tests with the single weights $w_1 \equiv 1$ and $w_2(x) = 1 - 2x$ as well as the combination of them. Additionally, we perform a Cox- and Aalen-regression, respectively, with and without including the interaction effects.

TABLE 2 *p*-Values in % for the data example

				With Interaction		Without Interaction	
	Comb	LR	Cross	Cox	Aalen	Cox	Aalen
Farm	0.4	0.2	0.3	89.2	69.8	0.2	0.2
Sex	1.3	9.0	5.9	42.0	41.9	10.3	10.5
Interaction	80.1	53.6	68.4	37.5	54.6	–	–

Note: Comb: joint Wald-type test (Comb); LR: the singly weighted tests based on $w_1(t) = 1$; Cross: the singly weighted tests based on $w_2(t) = 1 - 2t$; Cox: the Cox-regression with and without interaction effect; Aalen: the Aalen-regression with and without interaction effect.

Ties have been broken by small random noise. The *p*-values of all methods are displayed in Figure 2. For decision making, we choose the level $\alpha = 5\%$.

All tests, except the two regression methods when an interaction effect is incorporated, reveal a significant effect of being raised on a farm. This confirms the findings of our simulation study, where the inclusion of (additional) interaction effects may lead to a significant power loss. However, only the joint Wald-type test (*p*-value: 1.3%) can detect a gender effect. Here, the diverse testing results, in particular, the difference between the combination approach and the cox regression, can be explained by performing some diagnostics regarding the assumption of proportional hazards. On the one hand, we applied the test of Grambsch and Therneau (1994) for the gender variable. The resulting *p*-values of 2.1% and 2.2% for the model with and without the interaction effect, respectively, indicate a violation of the proportional hazard assumption. Moreover, plots of kernel-based estimators for the hazard rates provided in the supplement show a clear violation of the proportional hazards assumption and crossings of the hazard rates for the two different genders. Altogether this underpins the strength of the combination approach that is independent of a specific model assumption. Both findings, that is, a main effect of the two factors, are in line with a naive graphical comparison in Figure 3. Finally, no test finds a significant interaction effect.

7 | DISCUSSION AND OUTLOOK

We proposed new inference methods for general factorial designs with right-censored time-to-event data. Compared to existing regression methods, we utilize a flexible nonparametric framework that works without any restrictive model assumptions. The basic ingredients of the new methods are extensions of weighted log-rank tests to the factorial design setup. In the two-sample situation, weighted log-rank tests are known to be optimal (Fleming and Harrington, 1991; Gill, 1980) for specific hazard alternatives. However, the weight needs to be calibrated to the typically unknown shape of the alternative and a wrong guess of the weight can lead to a very poor power performance. To address this tricky question of weight choice and to avoid blind guessing, Brendel et al. (2014), Ditzhaus and Pauly (2019), and Ditzhaus and Friedrich (2020) followed a combination approach of different weighted log-rank tests for the two-sample scenario. We extended this approach to general factorial survival designs. The advantage of the combination idea is that the respective tests do not have only a reasonable power in the directions of the chosen weights but covers even a larger alternative space, that is, more alternative directions. This can be seen in the simulation section and in the illustrative data analysis, where the combination approach (consisting of the log-rank weight and a crossing weight) possesses larger power than the singly weighted version with the crossing weight in most of the simulated settings with crossing hazard curves. The reason is that the crossing weight was not the optimal weight for these crossing alternatives and a shifted weight would have been a better guess. The advantage of the combination approach now is that all linear combinations of the prechosen weights are implicitly included. This particularly holds for the desired shifted weight. A more formal description on this issue is given in the Supporting Information. In the paper, we prove that our methods are asymptotically correct.

To facilitate application for small or moderate-sized studies, we suggested a permutation procedure and proved its theoretical validity under heterogeneous settings. Simulation studies for two-way designs confirm the resulting procedures' satisfactory type-I-error control and power. In the Supporting Information, these studies are complemented by additional simulations for one- and four-way designs. These results confirm the findings from the two-way situation: the combination approach is more flexible than the Cox- and Aalen-regression and leads to significantly higher power values under crossing hazard alternatives, while having reasonable power under proportional hazards. Besides, it turned out that including too many interaction effects into the two regression models lead to a significant loss in power, for example, including three-way interactions when testing for no two-way interaction. For the one-way setup, the simulations additionally illustrate that the proposed combination approach can compete with recently proposed testing methods (Qiu and

Sheng, 2008; Chen et al., 2017; Gorfine et al., 2020) that were explicitly designed to handle nonproportional hazard alternatives.

The resulting inference toolbox is coined CASANOVA abbreviating the presented cumulative Aalen survival analysis-of-variance approach. To bring the presented procedures into statistical practice, the authors implemented the combination test into the R-package *GFDsurv* available on CRAN, which contains additionally survival methods for factorial designs based on medians (Ditzhaus et al., 2021) and the concordance measure (Dobler and Pauly, 2020). Beside the implementation, the authors currently work on convincing medical doctors and epidemiologists to apply the methods in biostatistical cooperations.

We remark some possible extensions: The proofs for the asymptotic Wald-type tests rely on the fact that $N = (N_1, \dots, N_k)$ fulfills the multiplicative intensity model of Aalen (1978). More complex filtering mechanisms, for example, left truncation or certain interval censorings, can be endowed in the same methodology (Andersen et al., 1993). In fact, the results from Sections 3.1 and 3.2 remain true and the Wald-type statistic with a χ^2_f approximation can be applied. However, the permutation technique's validity is unclear. That is why multiplier resampling as investigated in Lin (1997); Dobler et al. (2017, 2019) would be our first choice to approximate the asymptotic χ^2_f-quantile in these general cases. Due to the permutation approach's beneficial properties for our setup, however, we omit this here and postpone corresponding investigations to future research, where we plan to study factorial designs in more complex multistate models (Bluhmki et al., 2018).

ACKNOWLEDGMENTS

The authors are grateful to the editor, the associate editor, and the two referees for their comments that substantially improved the paper's quality. Marc Ditzhaus and Markus Pauly gratefully acknowledge support from the *Deutsche Forschungsgemeinschaft* (grant no. PA-2409 5-1) and the computing time provided on the Linux HPC cluster at TU Dortmund (LiDO3), partially funded by the *Deutsche Forschungsgemeinschaft* as project 271512359. Furthermore, the authors thank Philipp Steinhauer for computational assistance.

DATA AVAILABILITY STATEMENT

The data that support the findings in this paper are available on request from the author Jon Genuneit (email: jon.genuneit@medizin.uni-leipzig.de). The data are not publicly available due to privacy or ethical restrictions.

ORCID

Marc Ditzhaus https://orcid.org/0000-0001-9235-1905
Jon Genuneit https://orcid.org/0000-0001-5764-1528

REFERENCES

Aalen, O. (1978) Nonparametric inference for a family of counting processes. *Annals of Statistics*, 6, 701–726.

Aalen, O. (1980) A model for nonparametric regression analysis of counting processes. In: *Mathematical Statistics and Probability Theory (Proc. Sixth Internat. Conf., Wisła, 1978)*, volume 2 of *Lecture Notes Statistics*. New York-Berlin: Springer, pp. 1–25.

Akritas, M. & Brunner, E. (1997) Nonparametric methods for factorial designs with censored data. *Journal of the American Statistical Association*, 92, 568–576.

Andersen, P., Borgan, O., Gill, R. D. & Keiding, N. (1993) *Statistics Models Based Counting Processes*. Springer Series in Statistics. New York: Springer-Verlag.

Bathke, A., Kim, M.-O. & Zhou, M. (2009) Combined multiple testing by censored empirical likelihood. *Journal of Statistical Planning and Inference*, 139, 814–827.

Bluhmki, T., Schmoor, C., Dobler, D., Pauly, M., Finke, J., Schumacher, M. et al. (2018) A wild bootstrap approach for the Aalen–Johansen estimator. *Biometrics*, 74, 977–985.

Bouliotis, G. & Billingham, L. (2011) Crossing survival curves: alternatives to the log-rank test. *Trials*, 12, A137.

Brendel, M., Janssen, A., Mayer, C.-D. & Pauly, M. (2014) Weighted logrank permutation tests for randomly right censored life science data. *Scandinavian Journal of Statistics*, 41, 742–761.

Chen, Z., Huang, H. & Qiu, P. (2016) Comparison of multiple hazard rate functions. *Biometrics*, 72, 39–45.

Chen, Z., Huang, H. & Qiu, P. (2017) An improved two-stage procedure to compare hazard curves. *Journal of Statistical Computation and Simulation*, 87, 1877–1886.

Chung, E. & Romano, J. (2013) Exact and asymptotically robust permutation tests. *Annals of Statistics*, 41, 484–507.

Cox, D. (1972) Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34, 187–202.

Ditzhaus, M., Dobler, D. & Pauly, M. (2021) Inferring median survival differences in general factorial designs via permutation tests. *Statistical Methods in Medical Research*, 30, 875–891.

Ditzhaus, M. & Friedrich, S. (2020) More powerful logrank permutation tests for two-sample survival data. *Journal of Statistical Computation and Simulation*, 90, 2209–2227.

Ditzhaus, M. & Janssen, A. (2020) Bootstrap and permutation rank tests for proportional hazards under right censoring. *Lifetime Data Analysis*, 26, 493–517.

Ditzhaus, M. & Pauly, M. (2019) Wild bootstrap logrank tests with broader power functions for testing superiority. *Computational Statistics & Data Analysis*, 136, 1–11.

Ditzhaus, M., Yu, M. & Xu, J. (2021) Studentized permutation method for comparing restricted mean survival times with small sample from randomized trials. *arXiv preprint arXiv:2102.10186*.

Dobler, D., Beyersmann, J. & Pauly, M. (2017) Non-strange weird resampling for complex survival data. *Biometrika*, 104, 699–711.

Dobler, D. & Pauly, M. (2014) Bootstrapping Aalen–Johansen processes for competing risks: handicaps, solutions, and limitations. *Electronic Journal of Statistics*, 8, 2779–2803.

Dobler, D. & Pauly, M. (2020) Factorial analyses of treatment effects under independent right-censoring. *Statistical Methods in Medical Research*, 29, 325–343.

Dobler, D., Pauly, M. & Scheike, T. (2019) Confidence bands for multiplicative hazards models: flexible resampling approaches. *Biometrics*, 75, 906–916.

Fernández, T. & Rivera, N. (2020) A reproducing kernel Hilbert space log-rank test for the two-sample problem. *Scandinavian Journal of Statistics*, 1–49. https://doi.org/10.1111/sjos.12496

Fleming, T. & Harrington, D. (1991) *Counting processes survival analysis*. New York: John Wiley & Sons, Inc.

Gahrton, G., Iacobelli, S., Björkstrand, B., Hegenbart, U. et al. (2013) Autologous/reduced-intensity allogeneic stem cell transplantation vs autologous transplantation in multiple myeloma: long-term results of the EBMT-NMAM2000 study. *Blood*, 121, 5055–5063.

Genuneit, J. (2012) Exposure to farming environments in childhood and asthma and wheeze in rural populations: a systematic review with meta-analysis. *Pediatric Allergy and Immunology*, 23, 509–518.

Genuneit, J. (2014) Sex-specific development of asthma differs between farm and nonfarm children: a cohort study. *American Journal of Respiratory and Critical Care Medicine*, 190, 588–590.

Genuneit, J., Büchele, G., Waser, M., Kovacs, K., Debinska, A., Boznanski, A. et al. (2011) The GABRIEL advanced surveys: study design, participation and evaluation of bias. *Pediatric Allergy and Immunology* 25, 436–447.

Gill, R. (1980) *Censoring and stochastic integrals*, Volume 124 of Mathematical Centre Tracts. Amsterdam: Mathematisch Centrum.

Gorfine, M., Schlesinger, M. & Hsu, L. (2020) K-sample omnibus non-proportional hazards tests based on right-censored data. *Statistical Methods in Medical Research*, 29, 2830–2850.

Grambsch, P. & Therneau, T. (1994) Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81, 515–526.

Green, S. (2012) Factorial designs with time to event endpoints. In *Handbook of statistics in clinical oncology*, 3rd ed. Boca Raton: Chapman and Hall/CRC, pp. 201–210.

Harrington, D. & Fleming, T. (1982) A class of rank test procedures for censored survival data. *Biometrika*, 69, 553–566.

Jacobs, I., Menon, U., Ryan, A., Gentry-Maharaj, A., Burnell, M., Kalsi, J. et al. (2016) Ovarian cancer screening and mortality in the UK collaborative trial of ovarian cancer screening (UKCTOCS): a randomised controlled trial. *Lancet*, 387, 945–956.

Janssen, A. & Mayer, C.-D. (2001) Conditional studentized survival tests for randomly censored models. *Scandinavian Journal of Statistics*, 28, 283–293.

Janssen, A. & Pauls, T. (2003) How do bootstrap and permutation tests work? *Annals of Statistics*, 31, 768–806.

Juszczak, E., Altman, D. G., Hopewell, S. & Schulz, K. (2019) Reporting of multi-arm parallel-group randomized trials: extension of the consort 2010 statement. *JAMA*, 321, 1610–1620.

Kristiansen, I. (2012) PRM39 Survival curve convergences and crossing: a threat to validity of meta-analysis? *Value Health*, 15, A652.

Kurz, A., Fleischmann, E., Sessler, D., Buggy, D., Apfel, C., Akça, O. et al. (2015) Effects of supplemental oxygen and dexamethasone on surgical site infection: a factorial randomized trial. *British Journal of Anaesthesia*, 115, 434–443.

Lehmann, E. & Romano, J. (2006) *Testing statistical hypotheses*. New York: Springer.

Li, H., Han, D., Hou, Y., Chen, H. & Chen, Z. (2015) Statistical inference methods for two crossing survival curves: a comparison of methods. *PLoS One*, 10, e0116774.

Lin, D. (1997) Non-parametric inference for cumulative incidence functions in competing risks studies. *Statistics in Medicine*, 16, 901–910.

Liu, T., Ditzhaus, M. & Xu, J. (2020) A resampling-based test for two crossing survival curves. *Pharmaceutical Statistics*, 19, 399–409.

Liu, Y. & Yin, G. (2017) Partitioned log-rank tests for the overall homogeneity of hazard rate functions. *Lifetime Data Analysis*, 23, 400–425.

Lubsen, J. & Pocock, S. (1994) Factorial trials in cardiology: pros and cons. *European Heart Journal*, 15, 585–588.

Mick, R. & Chen, T.-T. (2015) Statistical challenges in the design of late-stage cancer immunotherapy studies. *Cancer Immunology Research*, 3, 1292–1298.

Neuhaus, G. (1993) Conditional rank tests for the two-sample problem under random censorship. *Annals of Statistics*, 21, 1760–1779.

Pauly, M., Brunner, E. & Konietschke, F. (2015) Asymptotic permutation tests in general factorial designs. *Journal of the Royal Statistical Society: Series B*, 77, 461–473.

Qiu, P. & Sheng, J. (2008) A two-stage procedure for comparing hazard rate functions. *Journal of the Royal Statistical Society: Series B (Statistics and Methodology)*, 70, 191–208.

R Core Team (2020) *R: a language environment statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rao, C. & Mitra, S. (1971) *Generalized inverse of matrices and applications*. New York-London-Sydney: John Wiley & Sons, Inc.

Scheike, T. & Zhang, M.-J. (2003) Extensions and applications of the Cox–Aalen survival model. *Biometrics*, 59, 1036–1045.

Smith, M., Ridgway, P., Catton, C., Cannell, A., O'Sullivan, B. et al. (2014) Combined management of retroperitoneal sarcoma with dose intensification radiotherapy and resection: long-term results of a prospective trial. *Radiotherapy Oncology*, 110, 165–171.

Su, Z. & Zhu, M. (2018) Is it time for the weighted log-rank test to play a more important role in confirmatory trials? *Contemporary Clinical Trials Communications*, 10, A1.

Wang, R., Lagakos, S. & Gray, R. (2010) Testing and interval estimation for two-sample survival comparisons with small sample sizes and unequal censoring. *Biostatistics*, 11, 676–692.

Zein, J. G. & Erzurum, S. C. (2015) Asthma is different in women. *Current Allergy and Asthma Reports*, 15, 1–10.

SUPPORTING INFORMATION

Web Appendices, Tables, Figures, and the R-package referenced in Sections 2, 3, 5 and 7 are available with this paper at the Biometrics website on Wiley Online Library.

How to cite this article: Ditzhaus, M., Genuneit, J., Janssen, A. & Pauly, M. (2023) CASANOVA: Permutation inference in factorial survival designs. *Biometrics*, 79, 203–215. https://doi.org/10.1111/biom.13575