Studentized permutation method for comparing two restricted mean survival times with small sample from randomized trials

Marc Ditzhaus¹,² | Menggang Yu³ | Jin Xu⁴

¹Department of Statistics, TU Dortmund University, Dortmund, Germany
²Department of Mathematics, Otto-von-Guericke University, Magdeburg, Germany
³Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, Wisconsin, USA
⁴Key Laboratory of Advanced Theory and Application in Statistics and Data Science – MOE and School of Statistics, East China Normal University, Shanghai, China

Correspondence
Marc Ditzhaus, Institute of Mathematical Stochastics, Otto-von-Guericke University, Postfach 4120, 39106 Magdeburg, Germany.
Email: marc.ditzhaus@ovgu.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Numbers: 271512359, PA-2409 5-1; National Cancer Institute, Grant/Award Number: P50DE026787; National Institute of Dental and Craniofacial Research

Recent observations, especially in cancer immunotherapy clinical trials with time-to-event outcomes, show that the commonly used proportional hazard assumption is often not justifiable, hampering an appropriate analysis of the data by hazard ratios. An attractive alternative advocated is given by the restricted mean survival time (RMST), which does not rely on any model assumption and can always be interpreted intuitively. Since methods for the RMST based on asymptotic theory suffer from inflated type-I error under small sample sizes, a permutation test was proposed recently leading to more convincing results in simulations. However, classical permutation strategies require an exchangeable data setup between comparison groups which may be limiting in practice. Besides, it is not possible to invert related testing procedures to obtain valid confidence intervals, which can provide more in-depth information. In this paper, we address these limitations by proposing a studentized permutation test as well as respective permutation-based confidence intervals. In an extensive simulation study, we demonstrate the advantage of our new method, especially in situations with relatively small sample sizes and unbalanced groups. Finally, we illustrate the application of the proposed method by re-analyzing data from a recent lung cancer clinical trial.

KEYWORDS
hazard ratio, permutation methods, restricted mean survival time, survival analysis, time-to-event outcomes

1 | INTRODUCTION

While the log-rank test and hazard ratios were the gold standard in time-to-event analysis for a long time, there is a recent trend toward alternative methods not relying on the proportional hazard assumption. The reason for this change is recently observed violations of the proportional hazard assumption in real data. For example, Trinquart et al¹ analyzed 54 phase III oncology clinical trials from five leading journals, and in 13 (24%) of them the proportional hazard assumption could be rejected significantly. Especially in immunotherapy trials, a delayed treatment effect often leads to a violation of the proportional hazard assumption²,³ and suchlike could also be observed when comparing bone marrow transplant and chemotherapy for hematologic malignancies.⁴,⁵ More classical and known effect sizes as landmark...
survival\textsuperscript{6} and the median survival time\textsuperscript{7–9} provide rather a snapshot for a time point than information about the complete Kaplan-Meier curves. This may be one of the reasons why the restricted mean survival time (RMST)\textsuperscript{10–16} the integral of the Kaplan-Meier-curve over a clinically relevant time window, gets more and more attention lately. Since this summary measure is “arguably more helpful for clinical decision-making and more easily understood by patients,”\textsuperscript{17} methods based on it “should be routinely reported in randomized trials with time-to-event outcomes.”\textsuperscript{1}

Recently, Horiguchi and Uno\textsuperscript{18} pointed out that “there is a notable inflation of the type-I error rate” under small sample sizes when the asymptotic methods are used for two-sample comparisons in terms of the RMST. To overcome this problem, they suggested a permutation approach showing a significant improvement regarding the type-I error control in simulations. For small sample sizes, permutation methods are popular tools since they guarantee exact testing procedures\textsuperscript{19,20} for exchangeable data. In the right-censoring survival set-up, as considered for this paper, the exchangeability of the data translates into equal survival as well as equal censoring distributions in the two groups. However, the censoring distributions may differ between comparison groups. In addition, confidence intervals for the quantity of interest, here the difference or ratio of the RMSTs, cannot be derived as Horiguchi and Uno\textsuperscript{18} mentioned: “Further research to develop methods for constructing confidence intervals for RMST difference with a small sample data is warranted.” Therefore we propose another strategy known as studentized permutation which was already successfully applied to different two-sample survival and nonsurvival settings,\textsuperscript{21–25} one-way layouts\textsuperscript{8,26} and, more recently, even to factorial designs.\textsuperscript{9,27–32} Using the studentized permutation strategy we are able to solve the existing issues and, moreover, preserve the finite exactness under exchangeability. This is justified in the present paper on a theoretical level employing the empirical process theory\textsuperscript{33} as well as by an extensive simulation study. In contrast to the usual martingale argumentation, by empirical process theory\textsuperscript{33} we can handle even tied data, e.g. survival times rounded to days, months, etc.

The paper is organized as follows. First, our methodology is presented in Section 2. Therein, we explain in detail why the permutation test of of Horiguchi and Uno\textsuperscript{18} may fail for general nonexchangeable data and how studentization solves for that. Moreover, permutation-based confidence intervals for the difference and the ratio of RMSTs are presented. To empirically assess the performances of the proposed test and confidence interval, we conducted an extensive simulation study comparing the asymptotic and the two permutation methods in Section 3. Their applications are illustrated by analyzing data from a recent lung cancer trial in Section 4. Finally, we give some final remarks and discuss possible future extensions in Section 5. All proofs and some additional simulation results are given in the Supplement.

## 2 | METHODOLOGY

We consider the two-sample survival setup given by mutually independent survival and censoring times

$$T_{ij} \sim S_i, \quad C_{ij} \sim G_i \quad (i = 1, 2; \quad j = 1, \ldots, n_i),$$

respectively. Here, $S_i$ and $G_i$ denote the survival functions for the survival and censoring times of the $i$th group, respectively. Both are not necessarily continuous and ties in the data are explicitly allowed, for example, survival times rounded to days, months etc. Based on the right-censored event times $X_{ij} = \min(T_{ij}, C_{ij})$ and the censoring statuses $\delta_{ij} = 1\{X_{ij} = T_{ij}\}$, we would like to infer differences between the two groups in terms of their RMSTs

$$\mu_i = \int_0^\tau S_i(t) \ dt \quad (i = 1, 2),$$

over a prespecified time window $[0, \tau]$, which is practically relevant (eg, $\tau = 2$ years). Thereby, it needs to be guaranteed that the event times $X_i$ larger than $\tau$ are observable with a positive probability $P(X_{ij} \geq \tau) > 0$. In practice, a typical choice for $\tau$ is the end-of-study time. While $\tau$ is usually be chosen as a prespecified constant allowing a straight-forward interpretation of $\mu_i$, Tian et al.\textsuperscript{34} discuss an empirical choice of $\tau$, for example, the largest observed time, under appropriate regularity assumptions on the censoring distribution.

The RMST can be naturally estimated by plugging-in the Kaplan-Meier estimator $\hat{S}_i$;

$$\hat{\mu}_i = \int_0^\tau \hat{S}_i(t) \ dt \quad (i = 1, 2).$$

Asymptotic inference for this estimator relies on a normal approximation, which can be justified by martingale arguments\textsuperscript{35} combined with the continuous mapping theorem. However, this argumentation, as, for example, used by Zhao
et al for the RMST, relies on continuously distributed survival and censoring times. Adopting the empirical process technique instead, the same asymptotic result can be derived even when ties are present, see the Data S1 for a detailed verification. In fact, under the assumption of non-vanishing groups, that is, $n_i/n \to \kappa_i \in (0,1)$ as $n \to \infty$, which is supposed throughout the paper, we obtain

$$\sqrt{n} \left\{ (\hat{\mu}_1 - \hat{\mu}_2) - (\mu_1 - \mu_2) \right\} \overset{d}{\to} Z \sim N(0,\sigma^2), \quad \sigma^2 = \sigma_1^2 + \sigma_2^2. \tag{1}$$

Here, $\sigma_i^2$ denotes the asymptotic variance of $\sqrt{n}(\hat{\mu}_i - \mu_i)$ and is given by

$$\sigma_i^2 = \kappa_i^{-1} \int_0^\tau \left\{ \int_x^\tau S_i(t) \, dt \right\}^2 \frac{1}{\{1 - \Delta A_i(x)\} G_{i-}(x) S_{i-}(x)} \, dA_i(x) \quad (i = 1,2),$$

where $A_i = -\log(S_i)$ is the cumulative hazard rate function and $\Delta A_i(x) = A_i(x) - A_{i-}(x)$ is its increment in $x$. Moreover, $G_{i-}$, $S_{i-}$ and $A_{i-}$ denote the left-continuous versions of $G_i$, $S_i$ and $A_i$, respectively, for example, $G_{i-}(t) = P(C_i \geq t)$ (c.f. $G_i(t) = P(C_i > t)$). The variance can be estimated straightforwardly by replacing $S_i$, $G_i$ and $A_i$ by their respective Kaplan-Meier ($\hat{S}_i$, $\hat{G}_i$) and Nelson-Aalen ($\hat{A}_i$) estimators. In detail, $\hat{\sigma}^2 = \hat{\sigma}_1^2 + \hat{\sigma}_2^2$ and

$$\hat{\sigma}_i^2 = \frac{n}{n_i} \int_0^\tau \left\{ \int_x^\tau \hat{S}_i(t) \, dt \right\}^2 \frac{1}{\{1 - \Delta \hat{A}_i(x)\} \hat{S}_{i-}(x) \hat{G}_{i-}(x)} \, d\hat{A}_i(x). \tag{2}$$

Combining (1) and (2), we obtain an asymptotically valid test $\varphi = \mathbf{1}\{ |\hat{\mu}_1 - \hat{\mu}_2| / \hat{\sigma} > z_{1-\alpha/2} \}$ for the null hypothesis of equal RMSTs:

$$H_0 : \mu_1 = \mu_2.$$

Here, $z_{1-\alpha/2}$ denotes the $(1 - \alpha/2)$-quantile of a standard normal distribution. However, for small sample sizes, this test has an inflated type-I error control, as seen in Horiguchi and Uno and Section 3. To tackle this problem, Horiguchi and Uno proposed a permutation approach. In the next subsection, we discuss their permutation approach as well as its limitations and propose an improved permutation strategy, both hypothesis testing and confidence interval construction.

### 2.1 Unstudentized permutation test and its studentized version

Following the idea of exact permutation tests, Horiguchi and Uno recently proposed a permutation test for $H_0 : \mu_1 = \mu_2$, which we call the unstudentized test hereafter.

In detail, given the observed data $(X, \delta) \equiv \{(X_{ij}, \delta_y): i = 1, 2; j = 1, \ldots, n_i \}$, let $(X^\tau, \delta^\tau) \equiv \{(X_{ij}^\tau, \delta_y^\tau): i = 1, 2; j = 1, \ldots, n_i \}$ be its permuted version corresponding to a random intermixing of the treatment indicator. Note that the permutation is at the subject level and $(X_{ij}, \delta_y)$ are permuted in pairs. Horiguchi and Uno suggested using the permutation test $\varphi_{\text{HU}}^\tau = \mathbf{1}\{|\hat{\mu}_1 - \hat{\mu}_2| > q_{1-\alpha,\text{HU}}^\tau \}$ in case of small sample sizes, where $q_{1-\alpha,\text{HU}}^\tau$ is the $(1 - \alpha)$-quantile of the permutation distribution $t \mapsto P\{ |\hat{\mu}_1^\tau - \hat{\mu}_2^\tau| \leq t \mid (X, \delta) \}$ given the observed data $(X, \delta)$. Here, $\hat{\mu}_1^\tau$, $\hat{\mu}_2^\tau$ denote the permutation counterparts of the original estimators by replacing the data $(X, \delta)$ with a permuted sample $(X^\tau, \delta^\tau)$.

Such permutation tests are known to be finally exact, that is, the type-I error is controlled not only asymptotically but for every fixed sample size, under exchangeable data. Recall that, in the context of right-censored survival data, exchangeability translates into equal survival and censoring distributions between the groups, respectively, that is, $S_1 = S_2$ and $G_1 = G_2$. This is obviously a much stronger assumption on both the interested time-to-event outcome and the censoring distributions. In our context of RMST comparison, having potentially crossing survival curves in mind, the null hypothesis $H_0 : \mu_1 = \mu_2$ may be true despite $S_1 \neq S_2$ holds, as shown in Figure 1. Besides, the assumption of equal censoring distributions alone is also too restrictive and typically unwanted for survival methods. An additional disadvantage is that this unstudentized permutation strategy cannot be used to obtain valid confidence intervals because the fact $\mu_1 \neq \mu_2$ clearly violates the exchangeability assumption.

To address all these issues, we propose a studentized permutation test. To explain our idea, we need to understand first the asymptotic behavior of the permuted, unstudentized statistic, here $\hat{\mu}_1 - \hat{\mu}_2$, under nonexchangeable settings. For that
FIGURE 1 The survival functions of the 9 different settings from Section 3 under the null hypothesis as well as under the alternative $\mu_2 = 1.5 + \mu_1$ (see A and B) and the survival functions for the three different censoring scenarios (see C). (A) Equal survival distributions under the null hypothesis; (B) Different survival distributions under the null hypothesis; (C) Censoring distributions.
purpose, we introduce the pooled Kaplan-Meier estimator \( \hat{S} \) and the pooled Nelson-Aalen estimator \( \hat{A} \). In detail, let \( N(t) = \sum_{ij} \delta_i \mathbf{1}[X_{ij} \leq t] \) be the number of events up until \( t \) and \( Y(t) = \sum_{ij} \mathbf{1}[X_{ij} \geq t] \) be the number of individuals under risk at time \( t \). Moreover, let \( t_1, \ldots, t_d \in \mathbb{N} \), be the distinctive time points within \( X \). Then \( \hat{S}(t) = \prod_{i:t_i \leq t} [1 - \Delta N(t_i)/Y(t_i)] \) and \( \hat{A}(t) = \sum_{i:t_i \leq t} \Delta N(t_i)/Y(t_i) \). Now, define \( v(t) = \sum_{i=1}^d \int_{t_i}^{t} G_{i-}(t) \, \text{d} F_i(s) \), where \( F_i(t) = 1 - S_i \).

Combining the Glivenko-Cantelli Theorem and the continuous mapping theorem we obtain almost surely that \( \hat{S}(t) \) and \( \hat{A}(t) \) converge uniformly on \([0, t]\) to \( S(t) = \exp[-A(t)] \) and \( A(t) = \int_{t}^{t_1} 1/y(s) \, \text{d}v(s) \), respectively, see the Data S1 for more details. Having these additional notations at hand, we are now able to derive the asymptotic limit of the permuted, unstudentized statistic \( \hat{\mu}_2^*-\hat{\mu}_2^* \).

**Theorem 1.** Under \( H_0 : \mu_1 = \mu_2 \) as well as under \( H_1 : \mu_1 \neq \mu_2 \), the conditional distribution of the unstudentized permutation statistic \( \sqrt{n}(\hat{\mu}_1^* - \hat{\mu}_2^*) \) given the data \((X, \delta)\) converges to a centred normal distribution with limiting variance given by

\[
\sigma_{\text{perm}}^2 = \frac{1}{k_1k_2} \int_0^\infty \left\{ \int_{\infty}^\chi S(t) \, \text{d}t \right\}^2 \frac{1}{[1-\Delta A(x)]} \frac{\text{d}A(x)}{y(t)},
\]

that is, when \( \Phi \) denotes the standard normal distribution function we have

\[
\sup_{t \in \mathbb{R}} \left| P \left\{ \sqrt{n}(\hat{\mu}_1^* - \hat{\mu}_2^*) \leq t \mid (X, \delta) \right\} - \Phi \left( \frac{t}{\sigma_{\text{perm}}} \right) \right| \xrightarrow{p} 0 \ \text{as} \ n \to \infty.
\]

In the special case \( S_1 = S_2 \) and \( G_1 = G_2 \), the variances \( \sigma^2 \) in (1) and \( \sigma_{\text{perm}}^2 \) coincide. But, in general, they are different. Thus, applying the unstudentized permutation test for a nonexchangeable setting may lead to a systematic error, which is caused by a different variance of the permuted statistic. However, this can be solved by studentization, that is, by including an appropriate variance estimator in the original test statistic as well as in its permutation counterpart. In fact, it can be shown that the permutation counterpart \( \tilde{\sigma}^2 \) of the variance estimator \( \hat{\sigma}^2 \) converges, given the observed data, to the variance \( \sigma_{\text{perm}}^2 \) from Theorem 1. In other words, the inclusion of the variance estimator in the permutation step corrects the wrong variance. Consequently, we obtain Theorem 2.

**Theorem 2.** Under \( H_0 : \mu_1 = \mu_2 \) as well as under \( H_1 : \mu_1 \neq \mu_2 \), the conditional distribution of the studentized permutation statistic \( \sqrt{n}(\hat{\mu}_1^*-\hat{\mu}_2^*)/\hat{\sigma}^* \) given the data \((X, \delta)\) converges to a standard normal distribution, that is,

\[
\sup_{t \in \mathbb{R}} \left| P \left\{ \sqrt{n}(\hat{\mu}_1^*-\hat{\mu}_2^*)/\hat{\sigma}^* \leq t \mid (X, \delta) \right\} - \Phi(t) \right| \xrightarrow{p} 0 \ \text{as} \ n \to \infty.
\]

From Theorem 2 we obtain that the conditional distribution of \( \sqrt{n}(\hat{\mu}_1^*-\hat{\mu}_2^*)/\hat{\sigma}^* \) mimics the asymptotic distribution of \( \sqrt{n}(\hat{\mu}_1 - \hat{\mu}_2 - (\mu_1 - \mu_2)) / \hat{\sigma} \), namely \( \{ Z \} \) for \( Z \sim N(0, 1) \). This is even true under any alternative. Therefore, we are able to formulate asymptotically valid confidence intervals based on the permutation procedure. For that purpose, let \( q_{1-a}^* \) denote the \((1-a)\)-quantile of the conditional distribution \( t \mapsto P\{ \sqrt{n}(\hat{\mu}_1^*-\hat{\mu}_2^*)/\hat{\sigma}^* \leq t \mid (X, \delta) \} \). Then the studentized permutation test \( \varphi^* \) and the permutation-based confidence interval \( I^* \) for \( \mu_1 - \mu_2 \) are given by

\[
\varphi^* = 1 \left\{ \sqrt{n} \frac{|\hat{\mu}_1 - \hat{\mu}_2|}{\hat{\sigma}} > q_{1-a}^* \right\}, \quad I^* = \left[ \hat{\mu}_1 - \hat{\mu}_2 \pm \frac{n^{-1/2} \hat{\sigma} q_{1-a}^*}{2} \right].
\]

Combining (1), Theorem 2, as well as lemma 1 and theorem 7 of Janssen and Pauls, we can deduce that the conditional quantile \( q_{1-a}^* \) tends to \( q_{1-a/2} \). Consequently, the permutation test is indeed asymptotically valid and the confidence interval has asymptotic coverage of \( 1 - \alpha \). In the same manner, confidence intervals for the ratio can be derived by considering a log-transformation \( \log(\hat{\mu}_1) - \log(\hat{\mu}_2) \). Analogous to (1), it follows

\[
\sqrt{n} \left\{ \log(\hat{\mu}_1) - \log(\hat{\mu}_2) \right\} \to N(0, \sigma_{\text{rat}}^2), \quad \sigma_{\text{rat}}^2 = \frac{\sigma_1^2}{\mu_1^2} + \frac{\sigma_2^2}{\mu_2^2},
\]
The asymptotic variance can be estimated by 

\[ \hat{\delta}^2_{\text{rat}} = (\hat{\delta}_1^2/\hat{\mu}_1^2) + (\hat{\delta}_2^2/\hat{\mu}_2^2). \]

Thus, an asymptotically valid confidence interval for \( \mu_1/\mu_2 \) and its studentized permutation counterpart are given, respectively, by

\[
I_{\text{rat}} = \left[ \exp \left\{ \log(\hat{\mu}_1) - \log(\hat{\mu}_2) \pm n^{-1/2} \hat{\delta}_{\text{rat}} q_{1-a/2} \right\} \right],
\]

\[
I^*_{\text{rat}} = \left[ \exp \left\{ \log(\hat{\mu}_1) - \log(\hat{\mu}_2) \pm n^{-1/2} \hat{\delta}_{\text{rat}}^* q_{1-a/2} \right\} \right],
\]

where \( q_{1-a/2} \) denotes the \((1 - a)\)-quantile of the conditional distribution \( t \mapsto P(\sqrt{n} | \log(\hat{\mu}_1^2) - \log(\hat{\mu}_2^2)/\hat{\delta}_{\text{rat}} \leq t | \mathbf{X}, \delta) \).

Summing up the results, we get Theorem 3.

**Theorem 3.** (i) The permutation test \( I^* \) has asymptotic level \( \alpha \) for \( H_0 : \mu_1 = \mu_2 \) and is consistent for general alternatives \( H_1 : \mu_1 \neq \mu_2 \), i.e. \( E_{H_0}(I^*) \rightarrow \alpha \) and \( E_{H_1}(I^*) \rightarrow 1 \) as \( n \rightarrow \infty \). (ii) The permutation-based confidence intervals \( I^* \) and \( I^*_{\text{rat}} \) have asymptotic confidence level \( 1 - \alpha \), that is, \( P(\mu_1 - \mu_2 \in I^*) \rightarrow 1 - \alpha \) and \( P(\mu_1/\mu_2 \in I^*_{\text{rat}}) \rightarrow 1 - \alpha \) as \( n \rightarrow \infty \).

## 3 | SIMULATIONS

To complement our theoretical discussion from the previous section, we conducted an extensive simulation study to examine the performance of the permutation test as well as the permutation-based confidence intervals. For ease of presentation, we restricted ourselves to the difference of the RMSTs. Additional results for the ratio are deferred to the Supplement (Data S1).

### 3.1 | Setup

We considered nine different choices for the survival times distribution, where for the first six settings the survival distributions of the two groups coincide under \( H_0 \) and for the last three choices the curves cross.

- **S1** Exponential distributions and proportional hazard alternatives: \( T_{11} \sim \text{Exp}(0.2) \) and \( T_{21} \sim \text{Exp}(0.5) \).
- **S2** Exponential distributions and late departures: \( T_{11} \sim \text{Exp}(0.2) \) and \( T_{21} \) with piece-wise constant hazard function \( \alpha_2(t) = 0.2 \cdot 1(t \leq 2) + 0.01 \cdot 1(t > 2) \).
- **S3** Exponential distributions and early departures: \( T_{11} \sim \text{Exp}(0.2) \) and \( T_{21} \) with piece-wise constant hazard function \( \alpha_3(t) = 0.5 \cdot 1(t \leq 0.8) + 0.1 \cdot 1(t > 0.8) \).
- **S4** Weibull distributions and late departures: \( T_{11} \sim \text{Weib}(3,8) \) and \( T_{21} \sim \text{Weib}(0.5,3,8/0.5) \).
- **S5** Weibull distributions and proportional hazard alternatives: \( T_{11} \sim \text{Weib}(3,8) \) and \( T_{21} \sim \text{Weib}(0.5,8) \).
- **S6** Lognormal distributions with scale alternatives: \( T_{11} \sim \text{logN}(2,0.25) \) and \( T_{21} \sim \text{logN}(2,0.25) \).
- **S7** Exponential vs piecewise Exponential: \( T_{11} \sim \text{Exp}(0.2) \) and \( T_{21} \) with piece-wise constant hazard function \( \alpha_5(t) = 0.5 \cdot 1(t \leq 0.7) + 0.05 \cdot 1(t > 0.7) \).
- **S8** Weibull distributions with crossing curves and shape alternatives: \( T_{11} \sim \text{Weib}(3,8) \) and \( T_{21} \sim \text{Weib}(0.5,8,14) \).
- **S9** Weibull distributions with crossing curves and scale alternatives: \( T_{11} \sim \text{Weib}(3,8) \) and \( T_{21} \sim \text{Weib}(0.5,3,15) \). 

The parameters \( \lambda_{k, k} \) depend on the difference \( \delta = \mu_2 - \mu_1 \) of the RMSTs. For our simulations, we considered \( \delta = 0 \) for the settings under the null hypotheses and \( \delta \in \{0.5, 1, 1.5\} \) for the different alternative scenarios. Note that under the null hypothesis \( (\delta = 0) \) Scenarios S1 to S3 as well as S4 and S5 coincide, respectively. That is why just one of the respective scenarios was included in the simulation study whenever \( \delta = 0 \) was considered. For the censoring, we chose the following three censoring configurations:

- **C1** unequally Weibull distributed censoring (Weib, uneq): \( C_{11} \sim \text{Weib}(3,18) \) and \( C_{21} \sim \text{Weib}(0.5,40) \).
- **C2** equally uniformly distributed censoring (Unif, eq): \( C_{11} \sim \text{Unif}(0,25) \) and \( C_{21} \sim \text{Unif}(0,25) \).
- **C3** equally Weibull distributed censoring (Weib, eq): \( C_{11} \sim \text{Weib}(3,15) \) and \( C_{21} \sim \text{Weib}(3,15) \).

In Figure 1, the survival curves for \( \delta \in \{0, 1.5\} \) and the censoring distributions are illustrated. For all simulations, we studied one balanced \( n_{\text{bal}} = (15, 15) \) and two unbalanced, \( n_{\text{incr}} = (12, 18) \) and \( n_{\text{decr}} = (18, 12) \), sample size settings and, additionally, considered their multiples \( Kn_{\text{bal}}, Kn_{\text{incr}}, Kn_{\text{decr}} \) with \( K = 2, 4 \) for larger sample sizes.
For the type-I error and power comparisons, we included the asymptotic test, the studentized and unstudentized permutation tests as well as the recent proposal by Zhou\textsuperscript{38} based on empirical likelihood ratios. While we programmed the asymptotic and the studentized permutation test by ourselves, the evaluation of the unstudentized permutation test was carried out utilizing the R-package \textit{survRM2perm}\textsuperscript{39} and the one of empirical likelihood approach using the R-package \textit{emplik}\textsuperscript{40} following the code examples of Zhou\textsuperscript{38}. Horiguchi and Uno\textsuperscript{18} discussed extensively different strategies on tackling the problem of possibly inestimable Kaplan-Meier-estimators for permuted data sets. Here, the curves are called inestimable\textsuperscript{18} when at least for one group the largest observed time was censored and is smaller than \( \tau \). However, their numerical findings do not reveal a clear favorable method and all six studied strategies lead to comparable results. That is why we restricted ourselves here mainly to the simple horizontal extension of the Kaplan-Meier curves, which corresponds to Method 2 in their paper and R-package. In detail, we set \( \hat{S}_i^T(u) = \hat{S}_i^T(t) \) for all \( u \in [t, \tau] \) when \( \hat{S}_i^T \) was just estimable up to \( t < \tau \). As suggested by a referee, we performed an additional comparison of this method and the regenerating procedure of Horiguchi and Uno\textsuperscript{18}, that is, permute again when the permuted Kaplan-Meier curves are inestimable, corresponding to Method 1 in their paper and R-package. This strategy may be the first choice according to Horiguchi and Uno\textsuperscript{18} when computation time is not an issue. As expected and known for the unstudentized tests, the results between both strategies for the novel studentized permutation test are almost indistinguishable. Therefore, we mainly present the ones for the horizontal extension (Method 2) and just show the ones for the regenerating procedure (Method 1) when it does not affect the readability.

The unstudentized permutation method, which relies on the assumption of exchangeable data, cannot be used to derive confidence intervals. Consequently, just the asymptotic, studentized permutation and empirical likelihood ratio methods were included in the respective comparisons.

The simulations were conducted by means of the computing environment \textit{R},\textsuperscript{41} version 3.6.1, generating \( N_{\text{sim}} = 5000 \) simulation runs and \( N_{\text{res}} = 2000 \) resampling iterations for the permutation procedures. Analogous to Horiguchi and Uno\textsuperscript{18}, we regenerated the data whenever the Kaplan-Meier-estimator was inestimable. The nominal significance level was set to \( \alpha = 5\% \) and the endpoint of the time window was set to \( \tau = 10 \).

### 3.2 Results

The simulation results for the type-I error control are presented in Table 1. Therein, we just present a sketch of the results to improve the readability of the paper, namely we restrict to the settings (S6) and (S8) combined with the censoring settings (C1) and (C2). The remaining results are given in the Data S1. There it can be seen that the settings of equal survival functions (S1-S6) as well as the ones with different survival distributions but equal RMSTs (S7-S9) lead, in general, to the same, partially less extreme, conclusions, respectively. To judge the tests’ performance, we recall that the 95%-confidence interval for the estimated sizes based on \( N_{\text{sim}} = 5000 \) simulation runs equals \([4.4\%, 5.6\%]\) if the true type-I error coincides indeed with the nominal level \( \alpha = 5\% \). Having this at hand, it can readily be seen that the asymptotic approach leads to rather liberal decisions. The liberality is most pronounced under the small sample sizes (\( K = 1 \)) settings with values up to 9.4%. However, the empirical sizes come closer to the nominal level when the sample sizes increased and in the majority of the largest sample size cases (\( K = 4 \)) the empirical size was inside the confidence interval \([4.4\%, 5.6\%]\) or, at least, very close to it. These observations complement the findings of Horiguchi and Uno\textsuperscript{18}, who considered only exchangeable and balanced settings. Therefore, the asymptotic test cannot be recommended for rather small sample sizes.

The empirical likelihood ratio approach of Zhou\textsuperscript{38} leads, in general, to less liberal decisions than the asymptotic test. While liberality is clearly present for the small sample sizes (\( K = 1 \)) with values up to 7.2%, it vanishes almost completely already for the moderate settings (\( K = 2 \)). The results confirm the simulation results of Zhou\textsuperscript{18} that the empirical likelihood ratio test leads to a more accurate control than the asymptotic test, but also show that there is still room for improvement for rather small sample sizes.

Switching to the unstudentized permutation test, the findings are diverse. For equal survival distributions (S6), combined with equally distributed censoring (C2), the permutation keeps the nominal level very accurately. This observation is not surprising because these settings correspond to an exchangeable data situation. However, this is not true anymore when unequal censoring is considered instead. Here, the test exhibits a slight conservativeness for \( n = (12, 18) \) and a slight liberality for \( n = (18, 12), (72, 48) \). The empirical sizes under Scenarios S8 with crossing survival curves become even more unstable. For the equally distributed censoring setting, the unstudentized permutation test leads to rather liberal decisions for \( n = K \cdot (18, 12) \) with values up to 7.2% and quite conservative decisions for \( n = K \cdot (12, 18) \) with values reaching down to 3.3%. Moreover, the performance does not improve under these unbalance sample size settings even...
when the sample sizes increase. In contrast to these findings, the type-I error rate is again well preserved by the unstudentized permutation test under the balanced settings. However, this changes when we consider the censoring setting C1 with unequal censoring distributions. Here, the test exhibits a rather liberality with values even up to 9.6%. The conservativeness under the other unbalanced setting, that is, \( n = K \cdot (18, 12) \), is now less pronounced but still present with values between 3.9% and 4.3%.

The overall unstable type-I error performance can be explained by the systematic error mentioned in Section 2.1, which is caused by the difference between the variance of the test statistic and its permutation counterpart. As motivated there, this can be fixed by studentization. The studentized permutation tests keep the type-I error rate in all settings inside the binomial confidence interval, except for one setting. Overall it leads to the most stable results under the null hypothesis and we recommend its application whenever the sample sizes are rather small (eg, \( n_1 + n_2 < 100 \)).

Due to limited space, the results of the power comparisons are deferred to the Data S1. Here, we just like to summarize some of the findings. The unstudentized permutation test leads partially to higher power values than the studentized permutation test with differences up to 6.9%. However, in some cases the reverse is true with differences up to 3.9%. A closer look at the related settings reveals that these differences are connected to the observed liberality and conservativeness, respectively, of the unstudentized permutation tests. Overall, the results, also for the asymptotic and the empirical likelihood ratio test, need to be taken with a pinch of salt, because only the studentized permutation test exhibited a generally convincing performance under the null hypotheses.

We finally turn to the performance of the confidence intervals. Here, we compare the asymptotic, the empirical likelihood ratio and the studentized permutation tests. For the permutation approach, we now differentiate between the regenerating strategy (Method 1) and the horizontal extension approach (Method 2). We summarized the results for all nine distributional choices (S1) to (S9), the three censoring distributions (C1) to (C3) and the four different choices for \( \delta \in \{0, 0.5, 1, 1.5\} \) in Figure 2, for each of the nine different sample sizes. In total, each boxplot summarizes the results of 99 different settings; recall that (S1) to (S3) and (S4) and (S5), respectively, coincide under \( \delta = 0 \) and, thus, only one of them is considered. It is apparent that the empirical coverage of the asymptotic test is liberal, similar to our findings regarding the type-I error control. The liberality or undercoverage is most pronounced for the small sample size cases (\( K = 1 \)) and becomes less pronounced when the sample sizes increase. But even for the largest sample size settings...
(K = 4), the median empirical coverage is closely around the lower border 94.4% of the binomial 95%-confidence interval [94.4%, 95.6%]. The results for the empirical likelihood ratio approach are similar but, again, the liberality/undercoverage is less pronounced than the one of the asymptotic strategy. While the undercoverage is clearly present for the smallest sample sizes (K = 1) and still partially present when doubling them (K = 2), the complete box of the empirical likelihood ratio test is within the binomial 95%-confidence interval [94.4%, 95.6%] for the largest sample sizes (K = 4) considered in the study. In contrast, the permutation-based confidence intervals based on both methods lead to more satisfactory results for all considered sample sizes and all boxes are clearly inside the 95%-confidence interval [94.4%, 95.6%], except a partial liberality/undercoverage in the setting \( n = (18, 12) \). To judge the exception, it should be noted that the undercoverage of the other two approaches is most pronounced also for \( n = (18, 12) \). These findings can be explained by the unfavorable combination of the following three circumstances: (i) the sample size of the second group is very small, (ii) the unequal censoring scenario leads to higher censoring in the second group, (iii) the alternatives \( \delta = \mu_2 - \mu_1 \in \{0.5, 1.1, 1.5\} \) lead to higher survival times in the second group, implying more observations outside the window \([0, \tau]\) of the analysis and more censored data. For an in-depth inspection of the coverage results, we refer to the Data S1, where the detailed results are presented in tables. There it is apparent that the undercoverage of the permutation tests for \( n = (18, 12) \) mainly appears in the settings (S8) and (S9), and is slightly more pronounced in the latter with values reaching down to 91.6%. Increasing the sample sizes leads to an improvement but, at least, for \( n = (36, 24) \) a systematic undercoverage is still present. In the small balanced sample size settings \( n = (15, 15) \), the undercoverage can be observed in the settings (S8) and (S9) as well. However, it is less pronounced and vanishes when the sample sizes are doubled. For the remaining unbalanced sample size scenario \( n = (12, 18) \), the outliers are still in a rather acceptable range and we could not recognize any systematic pattern. Since they are just a handful, these values may also be explained by simulation variability.

In summary, we can only recommend the studentized permutation test and the corresponding permutation-based confidence intervals for the quantity \( \mu_1 - \mu_2 \) for small sample sizes, as it leads to the most accurate type-I error and coverage

---

**FIGURE 2** Coverage in % (nominal level \( \alpha = 5\% \)) of the confidence intervals based on the asymptotic approximation (Asy), the empirical likelihood ratio (ELR) and the studentized permutation approach with the regenerating strategy (Perm I) as well as with the horizontal extension of the permuted Kaplan-Meier curves (Perm II), respectively. The dashed, horizontal lines represent the binomial 95%-confidence interval [94.4%, 95.6%].
TABLE 2 Type-I error rates and power values in % (nominal level \( \alpha = 5\% \))

| Cens. | K | KONP1 | KONP2 | LR | mdir | RMST | KONP1 | KONP2 | LR | mdir | RMST |
|-------|---|-------|-------|----|------|------|-------|-------|----|------|------|
| S6: Lognormal distributions | | | | | | | | | | | |
| un. W. | 1 | 3.5 | 4.4 | 5.9 | 4.7 | 5.5 | 22.7 | 32.0 | 38.4 | 24.7 | 40.3 |
| | 2 | 4.8 | 5.1 | 5.8 | 5.1 | 4.7 | 48.7 | 60.7 | 64.6 | 55.0 | 67.9 |
| | 4 | 4.3 | 4.8 | 5.4 | 5.0 | 5.3 | 82.2 | 90.4 | 91.4 | 88.9 | 93.6 |
| eq. U. | 1 | 3.8 | 4.7 | 6.0 | 4.6 | 4.9 | 22.0 | 32.3 | 38.4 | 26.9 | 36.6 |
| | 2 | 4.2 | 4.7 | 5.9 | 4.7 | 4.7 | 44.9 | 57.8 | 64.4 | 54.9 | 66.1 |
| | 4 | 4.7 | 5.1 | 5.4 | 5.3 | 4.8 | 76.9 | 86.9 | 90.7 | 86.8 | 91.7 |
| eq. W. | 1 | 5.0 | 5.7 | 6.0 | 5.0 | 5.1 | 30.0 | 37.9 | 44.8 | 33.3 | 44.3 |
| | 2 | 4.5 | 4.7 | 5.1 | 4.4 | 4.7 | 56.8 | 67.3 | 72.5 | 64.9 | 73.4 |
| | 4 | 5.4 | 5.4 | 5.0 | 4.4 | 4.9 | 86.3 | 92.6 | 95.5 | 93.1 | 96.3 |
| S8: Weibull distributions with crossing curves and shape alternatives | | | | | | | | | | | |
| un. W. | 1 | 70.5 | 68.5 | 33.1 | 74.9 | 5.5 | 58.4 | 69.7 | 72.7 | 67.8 | 35.5 |
| | 2 | 97.7 | 97.3 | 59.4 | 99.0 | 5.4 | 93.1 | 96.7 | 96.3 | 97.1 | 58.5 |
| | 4 | 100.0 | 100.0 | 89.4 | 100.0 | 5.0 | 99.7 | 100.0 | 100.0 | 100.0 | 87.2 |
| eq. U. | 1 | 58.5 | 55.1 | 21.7 | 72.8 | 5.4 | 53.9 | 65.3 | 66.4 | 63.4 | 33.6 |
| | 2 | 93.7 | 92.3 | 44.8 | 98.3 | 5.3 | 90.2 | 95.0 | 93.9 | 95.2 | 57.7 |
| | 4 | 100.0 | 100.0 | 74.7 | 100.0 | 5.2 | 99.8 | 99.9 | 99.9 | 100.0 | 86.7 |
| eq. W. | 1 | 66.6 | 64.2 | 32.7 | 81.0 | 5.0 | 68.9 | 76.8 | 77.0 | 74.3 | 39.7 |
| | 2 | 96.6 | 96.1 | 60.8 | 99.3 | 4.7 | 97.0 | 98.7 | 97.7 | 98.5 | 66.5 |
| | 4 | 100.0 | 100.0 | 89.9 | 100.0 | 5.3 | 100.0 | 100.0 | 100.0 | 100.0 | 90.4 |

Note: The values inside the binomial confidence interval [4.4\%, 5.6\%] under \( H_0 \) are printed bold.

Abbreviations: eq. U., equal uniform censoring; eq. W., equal Weibull censoring; KONP1, chi-squared test of Gorffine et al; KONP2, Cauchy-combination test of Gorffine et al; LR, log-rank test; mdir, combination test of Ditzhaus and Friedrich; RMST, studentized permutation test for the RMST (Method 2); un. W., unequal Weibull censoring.

control, respectively. Moreover, it can compete in terms of power with the other strategies whenever a comparison is fair and not influenced by liberal decisions under the null hypothesis.

3.3 Comparison to novel two-sample tests

To complement the previous simulation study, we additionally compared the proposed studentized permutation test for the RMST with recent novel two-sample tests which are robust against the assumption of proportional hazards. In particular, we included the omnibus permutation strategy of Gorffine et al based on sample space partitioning, which is implemented in the R package KONPsurv. From the recommendation of the authors, we use their chi-squared type method (KONP1) in our comparison. Furthermore, we compare with a Cauchy-combination test incorporated in the R package mdir.logrank. Finally, we compare with the combination approach of differently weighted log-rank tests, also known as the multidirection log-rank test (mdir) and implemented in the R-package mdir.logrank.

For the simulation study, we followed the setup described in Section 2 and chose, in particular, \( N_{res} = 2000 \) for all involved permutation methods. A representative part of the results under the null hypothesis \( H_0 : \mu_1 = \mu_2 \) and an alternative \( H_1 : \mu_2 = \mu_1 + 1.5 \) are presented in Table 2 for the setting (S6) with equal survival curves under \( H_0 \) and the setting (S8) with crossing survival curves under \( H_0 \). The results for the other settings are deferred to Data S1. We also included results from the classical log-rank tests as bench mark.
Starting with (S6) under $H_0$, we can observe an overall satisfactory type-I error control of the considered procedures with a slightly conservative behavior of KONP1 and a slightly liberal type-I error control of the log-rank test, both mainly for small sample sizes. Similar observations can be made in the settings (S1) to (S5) based on results in Data S1. Turning to the power, the RMST approach and the log-rank test yield the highest values. For small sample sizes ($K = 1$), the detection rate of KONP2 is higher than that of mdir while this difference becomes smaller and even vanishes for larger sizes. We chose here on purpose the setting (S6) with the best results in favor of the RMST. In the settings (S2) and (S5) with late differences, the RMST yields significantly less power values than the competitors while mdir and KONP2 show the most promising results. Switching to the setting (S3) with early departures or (S1) and (S5) with proportional hazard alternatives, the differences in terms of power between the RMST procedure and the competitors become much smaller. Note that the gain in power by the novel tests (i.e., KONP1, KONP2, and mdir) is to be expected as these settings focus on a narrower or simpler null $H_{0,S} : S_1(t) = S_2(t)$ for $t \in [0, \tau]$.

Now consider the settings (S7) to (S9) where the survival times’ distributions differ but $H_0 : \mu_1 = \mu_2$ is true. In these settings, the competitors (KONP1, KONP2, LR, mdir) all had excessive type-I error rates.

4 | REAL DATA EXAMPLE

To illustrate the presented permutation-based methods, we reconsider the data analysis of Hellmann et al.\textsuperscript{42} who compared the combination treatment of nivolumab plus ipilimumab with chemotherapy among 299 patients with nonsmall-cell lung cancer. Their study focused on patients with a high tumor mutational burden, that is, at least 10 mutations per megabase. And the study endpoint was progression-free survival. Since the present methods are designed for small sample sizes, we conduct a relevant subgroup analysis, which was also done by Hellmann et al.\textsuperscript{42} In detail, we restrict to the patients having PD-L1 (tumor programmed death ligand 1) expression of at least 1%. Based on the published Kaplan-Meier curves in Hellmann et al\textsuperscript{42} and some additional information therein, for example, the risk table, we reconstructed the individual patient data following the procedure of Guyot et al.\textsuperscript{45} The respective Kaplan-Meier curves of the two treatment groups are displayed in Figure 3. Therein, we can observe a delayed treatment effect of nivolumab plus ipilimumab. Thus, the assumption of proportional hazards is questionable and can even be formally rejected by the well-established test of Grambsch and Therneau\textsuperscript{46} or the recent permutation-based proposal of Ditzhaus and Janssen\textsuperscript{25} (with 10,000 permutations). Both tests lead to a $p$-value less than 0.1%. That is why the original analysis of Hellmann et al\textsuperscript{42} using the hazard ratio (HR: 0.48 and 95\% CI [0.27, 0.85]) need to be considered carefully. The RMSTs offer the possibility to interpret the treatment effect easily beyond the Cox model. The $p$-values of the asymptotic, empirical likelihood

![Figure 3](image-url)  
Kaplan-Meier curves of the reconstructed data.
If you are interested in the difference in survival, you may consider the point estimates and confidence intervals for the ratio of RMSTs. Table 4 shows these estimates for the combination nivolumab plus ipilimumab compared to the chemotherapy. The point estimates and confidence intervals are based on 5000 permutations.

For $\tau = 15$ and $\tau = 18$, the results confirm the findings of Hellmann et al.\cite{9} that the combination nivolumab plus ipilimumab improves the progression-free time compared to the chemotherapy. The point estimates and confidence intervals in Table 4 help to quantify the improvement and can be interpreted easily. For example, the combination treatment leads on average to a longer progression-free time of 4.02 ± 3.05 months (95% confidence based on 5000 permutations) compared to the chemotherapy over the first 1.5 years.

It can be seen that the asymptotic approach leads to smaller $P$-values and narrower confidence intervals than its permutation counterpart. Secondly, the unstudentized permutation test and the empirical likelihood ratio test lead to more comparable $P$-values than the asymptotic approach. As pointed out in Section 3, the results of all three tests (asymptotic, empirical likelihood ratio, unstudentized permutation) need to be considered carefully due to potential inflation of the type-I error, especially for small and unbalanced sample sizes.

## 5 Discussion and Remarks

In the last years, the RMST became an important part of the statistical toolbox for survival data. Various researchers\cite{2,9,10} advise to use it, at least, as a complementary summary statistic, especially when the assumption of proportional hazards is in doubt. As raised by Horiguchi and Uno,\cite{11} the type-I error rate of related asymptotic methods is inflated for small
sample sizes. Their proposed permutation procedure as well as their detailed discussion of how to deal with inestimable Kaplan-Meier curves of the permuted data was an important step to solve that problem. However, their test’s application is limited to exchangeable data settings and, in particular, to equal survival and censoring distributions, respectively.

In this paper, we explained how studentization can tackle these limitations. For the present survival two-sample comparison, it allows us to apply permutation tests even in nonexchangeable data situations, that is, for different survival and/or censoring distributions, as well as to formulate corresponding confidence intervals for the quantities $\mu_1 - \mu_2$ and $\mu_1 / \mu_2$ of interest. Moreover, the finite control of the type-I error under exchangeability, which was the initial motivation for permutation tests, is not affected by the studentization strategy. Compared to their asymptotic counterparts, studentized permutation tests usually show a satisfactory type-I error control even for small sample sizes, as seen in Section 3.

Our framework can be extended in various directions, for example, to more complex censoring mechanisms such as competing risks and to subsets of RMST such as window mean life. In addition, more general study designs may be part of future research. For that purpose, we can follow Dobler and Pauly and Ditzhaus et al, who recently discussed permutation-based inference for the concordance measure and median survival times, respectively, or the nonparametric approach of Ditzhaus et al in the general context of factorial designs. Sample size determination can also be developed, in parallel to the asymptotic test based results.

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 was funded by the Deutsche Forschungsgemeinschaft (grant no. PA-2409 5-1 and DI 2906/1-2). Menggang Yu was partly supported by the Specialized Program of Research Excellence (SPORE) program, through the NIH National Institute for Dental and Craniofacial Research (NIDCR) and National Cancer Institute (NCI), grant P50DE026787. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. Moreover, the authors gratefully acknowledge the computing time provided on the Linux HPC cluster at TU Dortmund (LiDO3), partially funded in the course of the Large-Scale Equipment Initiative by the Deutsche Forschungsgemeinschaft as project 271512359. Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST
The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT
Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID
Marc Ditzhaus https://orcid.org/0000-0001-9235-1905
Jin Xu https://orcid.org/0000-0003-2434-7183

REFERENCES
1. Trinquart L, Jacot J, Conner S, Porcher R. Comparison of treatment effects measured by the hazard ratio and by the ratio of restricted mean survival times in oncology randomized controlled trials. J Clin Oncol. 2016;34(15):1813-1819.
2. Mick T, Chen TT. Statistical challenges in the design of late-stage cancer immunotherapy studies. Cancer Immunol Res. 2015;3(12):1292-1298.
3. Alexander B, Schoenfeld J, Trippa L. Hazards of hazard ratios-deviations from model assumptions in immunotherapy. N Engl J Med. 2018;378(12):1158-1159.
4. Zittoun R, Mandelli F, Willemze R, et al. Autologous or allogeneic bone marrow transplantation compared with intensive chemotherapy in acute myelogenous leukemia. N Engl J Med. 1995;333(4):217-223.
5. Scott B, Pasquini M, Logan B, et al. Myeloablative versus reduced-intensity hematopoietic cell transplantation for acute myeloid leukemia and myelodysplastic syndromes. J Clin Oncol. 2017;35(11):1154.
45. Li H, Han D, Hou Y, Chen H, Chen Z. Statistical inference methods for two crossing survival curves: a comparison of methods. *PLoS One*. 2015;10(1):e0116774.
46. Pak K, Uno H, Kim DH, et al. Interpretability of cancer clinical trial results using restricted mean survival time as an alternative to the hazard ratio. *JAMA Oncol*. 2017;3(12):1692-1696.
47. Hellmann M, Ciuleanu TE, Pluzanski A, et al. Nivolumab plus ipilimumab in lung cancer with a high tumor mutational burden. *N Engl J Med*. 2018;378(22):2093-2104.
48. Guyot P, Ades A, Ouwens M, Welton N. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Med Res Methodol*. 2012;12(1):1-13.
49. Grambsch P, Therneau T. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*. 1994;81(3):515-526.
50. Murray S, Tsiatis A. Sequential methods for comparing years of life saved in the two-sample censored data problem. *Biometrics*. 1999;55(4):1085-1092.
51. Zhao L, Tian L, Claggett B, et al. Estimating treatment effect with clinical interpretation from a comparative clinical trial with an end point subject to competing risks. *JAMA Cardiol*. 2018;3(4):357-358.
52. Dobler D, Pauly M. Factorial analyses of treatment effects under independent right-censoring. *Stat Methods Med Res*. 2020;29(2):325-343.
53. Lyu J, Hou Y, Chen Z. The use of restricted mean time lost under competing risks data. *BMC Med Res Methodol*. 2020;20(1):1-11.
54. Ye T, Yu M. A robust approach to sample size calculation in cancer immunotherapy trials with delayed treatment effect. *Biometrics*. 2018;74(4):1292-1300.
55. Wang R, Lagakos S, Gray R. Testing and interval estimation for two-sample survival comparisons with small sample sizes and unequal censoring. *Biostatistics*. 2010;11(4):676-692.
56. Paukner M, Chappell R. Window mean life. *Stat Med*. 2021;40(25):5521-5533.

**SUPPORTING INFORMATION**

Additional supporting information can be found online in the Supporting Information section at the end of this article.

---

**How to cite this article:** Ditzhaus M, Yu M, Xu J. Studentized permutation method for comparing two restricted mean survival times with small sample from randomized trials. *Statistics in Medicine*. 2023;42(13):2226-2240. doi: 10.1002/sim.9720