RESEARCH ARTICLE

A nonparametric relative treatment effect for direct comparisons of censored paired survival outcomes

Dennis Dobler1,2,3 | Kathrin Möllenhoff4

1Department of Mathematics, Faculty of Science, Vrije Universiteit Amsterdam, Amsterdam, North-Holland, The Netherlands
2Department of Statistics, TU Dortmund University, Dortmund, North-Rhine Westphalia, Germany
3Research Center Trustworthy Data Science and Security, University Alliance Ruhr, Dortmund, North-Rhine Westphalia, Germany
4Institute of Medical Statistics and Computational Biology (IMSB), Faculty of Medicine, University of Cologne, Cologne, North-Rhine Westphalia, Germany

Correspondence
Dennis Dobler, TU Dortmund University, Joseph-von-Fraunhofer-Str. 25, 44227, Dortmund, Germany.
Email: dobler@statistik.tu-dortmund.de

A frequently addressed issue in clinical trials is the comparison of censored paired survival outcomes, for example, when individuals were matched based on their characteristics prior to the analysis. In this regard, a proper incorporation of the dependence structure of the paired censored outcomes is required and, up to now, appropriate methods are only rarely available in the literature. Moreover, existing methods are not motivated by the strive for insights by means of an easy-to-interpret parameter. Hence, we seek to develop a new estimand-driven method to compare the effectiveness of two treatments in the context of right-censored survival data with matched pairs. With the help of competing risks techniques, the so-called relative treatment effect is estimated. This estimand describes the probability that individuals under Treatment 1 have a longer lifetime than comparable individuals under Treatment 2. We derive hypothesis tests and confidence intervals based on a studentized version of the estimator, where resampling-based inference is established by means of a randomization method. In a simulation study, we demonstrate for numerous sample sizes and different amounts of censoring that the developed test exhibits a good power. Finally, we apply the methodology to a well-known benchmark data set from a trial with patients suffering from diabetic retinopathy.

KEYWORDS
estimand, matched pairs, nonparametric statistics, randomization, survival analysis

1 | INTRODUCTION

The comparison of censored paired survival outcomes is a frequently addressed issue in many clinical trials, for example, when individuals were matched prior to the analysis to assess a treatment effect excluding confounding factors. Such matched-pair analyses, where subjects from each treatment group are matched according to their characteristics, can substantially reduce bias and achieve a high level of evidence, even when performing a randomized controlled trial was not possible.1,3 Besides the analysis of matched pairs, paired survival outcomes occur in numerous other situations. One example is given by the Diabetic Retinopathy Study.4 In this trial, patients with “high-risk” diabetic retinopathy* are treated and the time until loss of visual acuity is observed. Each patient acts as its own control, as one eye was randomized to a laser photocoagulation, while the other eye received no treatment, resulting in paired survival outcomes.
All of the situations mentioned above require the proper incorporation of the dependence structure of the paired censored outcomes into the analysis, with a particular focus on the comparison of treatments. From a statistical point of view, in general, this problem can be translated to tests for the stochastic superiority of one distribution to another, which is actually a very classical problem. For two independent and fully observable samples, the $t$-test, the median test, and the Mann-Whitney-$U$ test are three well-known solutions to it. Likewise, in the case of two dependent samples, that is, if data consist of (matched) pairs, the $t$-test, the sign test, and the Wilcoxon signed rank test based on the pair-wise differences could be used. Extensions to the case of two dependent and censored samples are not obvious but some have been developed in the literature for a few decades. The developed test statistics are based on ranked absolute within-pair differences of the possibly censored survival times, a difference of counting processes, differences of efficient scores, ranking all censored observations separately from the uncensored ones, a combination of terms for different censoring and pair-wise ordering patterns, integrals of scores with respect to differences of the sample-specific Nelson-Aalen estimators for cumulative hazards, a combination of frailty-based log-rank tests, using further prioritized outcomes, that is, additional data, if the primary survival endpoint does not offer decisive pair-wise comparisons. Extensions for covariates were also developed. Reviews of methods for paired survival data including additional references to other approaches and discussions are also available.

While these existing approaches undoubtedly offer many good approaches for powerful statistical inference in the two-sample problem, an additional, easily interpretable quantification of the discrepancy between both samples is usually not available. One notable exception is the popular win ratio method which recently had been exploited through win odds in order to take ties into account in the inference method. The perhaps most straightforward approach for such a quantification is to compare the survival chances for both groups at a fixed time point. This would provide a very limited, yet easy-to-interpret summary. A more global impression of the difference between both samples could be obtained by the integration of survival curves over time, leading to the (restricted) mean survival time.

Instead, we will pursue an approach which is motivated by another estimand with a probabilistic interpretation: the relative treatment effect. It describes the probability that the lifetime under Treatment 1 is bigger than the lifetime under Treatment 2. Based on it, inferences can be drawn about the treatment efficacies. For example, if the probability is significantly bigger than 50%, one may call Treatment 1 more effective than Treatment 2. At the same time, it is a probability (and not, eg, a logarithmized hazard ratio), which might facilitate the communication of medical study results.

In the present paper, we will develop a nonparametric methodology with an emphasis on the following quality criteria:

1. begin the research with a clear formulation of an estimand of interest;
2. make only very few and weak assumptions for the method to work;
3. in particular, no continuity of the survival functions is needed, that is, instantaneous hazard rates need not exist;
4. develop hypothesis tests and confidence intervals about the estimand that are coherent with each other;
5. guaranteed large sample properties;
6. a good statistical reliability even for small samples, that is, good control of the type-I error rate and confidence level, as well as a good power and narrow confidence intervals, respectively.

All of these points are of crucial importance, in particular in the light of the ICH E9(R1) guidelines on estimands in trial analysis. At the beginning of Section A.5.1 therein, it is written that: “An estimand for the effect of treatment relative to a control will be estimated by comparing the outcomes in a group of subjects on the treatment to those in a similar group of subjects on the control. For a given estimand, an aligned method of analysis, or estimator, should be implemented that is able to provide an estimate on which reliable interpretation can be based. The method of analysis will also support calculation of confidence intervals and tests for statistical significance. An important consideration for whether an interpretable estimate will be available is the extent of assumptions that need to be made in the analysis.” These are precisely our goals.

These statements clearly put aspects of interpretability and reliability in the focus. Albeit the power of a test is of course a very important criterion for the method of choice, it is by far not the only one. For example, compared to log-rank tests, which are known to be powerful in many cases, it might still be preferable to use a test that is based on an easy-to-interpret estimand. This is all the more true when an application of the log-rank test does not make much sense, for example, in the case of crossing survival curves, because the corresponding hypotheses are stated in terms of the (in)equality or the ordering of the survival curves. In this connection, it should be pointed out that most existing methods for analyzing paired survival outcomes rely on strong assumptions such as the equality of censoring times for both members of a pair.
or the continuity of survival distributions. In this sense, the power of the test we will develop in this article is not the criterion of greatest importance although a powerful method is of course welcome.

This article is structured as follows. First, we will introduce the relative treatment effect and explain its estimation in Section 2. At the end of that section, we will relate the present approach to some others from the literature. Second, in Section 3, we will present a method to draw statistical inference based on a data re-randomization technique. Third, we will investigate the large sample properties of the new method and explore its small sample performance by means of a simulation study in Section 4. Therein, also the power of the test will be assessed in a comparison to some competitor methods. Next, in Section 5, we will apply the methodology to a well-known benchmark data set from the previously mentioned trial with patients suffering from diabetic retinopathy. We will conclude with a discussion in Section 6. The Supplementary Material contains all proofs, additional technical details, and additional simulation results.

2 PAIRED SURVIVAL AND THE RELATIVE TREATMENT EFFECT

2.1 Model and notation

We denote by \((T_1, T_2)\) a bivariate random vector on a probability space \((\Omega, \mathcal{A}, P)\). Each \(T_j (j = 1, 2)\) stands for a survival time of a patient who was randomized to receive treatment \(j\). For instance, the pair results from a matching of two individuals with a similar physiology; or, when medically plausible, from a patient who receives a treatment and, after the event occurred, receives another treatment after which another event time is measured. As a consequence, we generally assume \(T_1\) and \(T_2\) to be dependent. There exist many approaches for the estimation of the bivariate survival distribution of \((T_1, T_2)\) in the literature; see the paper by Pruitt for a comparison of six methods that are able to handle bivariate right-censored data; Dai et al developed an estimator under the more general assumption of bivariate left-truncated and right-censored data. We do not mean to give an exhaustive list of references in that direction and point to the references included in the two just mentioned papers. Instead of estimating the bivariate survival function, we aim at estimating a meaningful summary of it, that is, a treatment effect measure.

Let us now prepare the introduction of our estimand of interest for assessing the potential superiority of Treatment 1 over Treatment 2. Here, superiority means that Treatment 1 prolongs the survival times compared to Treatment 2 with a rather high probability; in Section 6, we will briefly discuss other criteria for superiority. To this end, we consider the following probability:

\[
\bar{\theta} = P(T_1 > T_2) + \frac{1}{2} P(T_1 = T_2).
\]

In words, \(\bar{\theta}\) is the probability that the subject who received the first treatment outlives the matched subject who received the second treatment. The second term is important to give equal credit to both treatments in the case of equal outcomes. Furthermore, we say that Treatment 1 is preferable if \(\bar{\theta} > 0.5\); other threshold values are also possible, depending on the medical context.

Because most studies have a finite time horizon, \(\bar{\theta}\) is typically not identifiable. Denote by \(\tau > 0\) such a maximum follow-up time. To take into account that a good fraction of survival times may exceed \(\tau\), we propose to focus on the following estimand:

\[
\theta = P(\min(T_1, \tau) > \min(T_2, \tau)) + \frac{1}{2} P(\min(T_1, \tau) = \min(T_2, \tau)).
\]

We call it the relative treatment effect. Throughout the article, we assume that \(\theta \in (0, 1)\), that is, no perfect superiority of one treatment over the other.

Brunner and Munzel introduced a similar concept (without the need to introduce \(\tau\)) for the case of two independent samples. In that context, Dobler and Pauly extended that method to the right-censored case with a terminal time \(\tau\). Based on the work of Brunner and Munzel, subsequent papers considered the case of two dependent but fully observable samples. In the present paper, we will develop extensions thereof under independent right-censoring regimes.

Remark 1 (Interpretation). Smaller values of \(\tau\) lead to higher probabilities of \(P(\min(T_1, \tau) = \min(T_2, \tau))\). In this sense, \(\theta\) has the tendency to be closer to 0.5 when \(\tau\) is small and, consequently, more pairs of observations
fall into the quadrant \([\tau, \infty) \times [\tau, \infty)\). For such small \(\tau\), the closer proximity to 0.5 reflects a bigger indecisiveness about the superiority of a treatment. A variant of the relative treatment effect that would not exhibit such tendency could be defined via restriction to the lower-left quadrant:

\[
\theta_{\vert \tau} = P(T_1 > T_2 \mid T_1 \leq \tau, T_2 \leq \tau) + \frac{1}{2} P(T_1 = T_2 \mid T_1 \leq \tau, T_2 \leq \tau).
\]

One potential disadvantage of \(\theta_{\vert \tau}\) is that it might strongly fluctuate for too small choices of \(\tau\). On the other hand, it might be of interest in practical application to report the relative or absolute contribution of \(P(T_1 \geq \tau, T_2 \geq \tau)\) to \(\theta\), which is part of the probability \(P(\min(T_1, \tau) = \min(T_2, \tau))\); see Figure 1 below.

**Remark 2 (Assumptions).** For technical reasons, we assume that \(P(T_1 > \tau, T_2 > \tau) > 0\), \(P(T_1 \neq T_2, \min(T_1, T_2) < \tau) > 0\), and \(P(T_1 = \tau) = P(T_2 = \tau) = 0\). The first assumption is due to mathematical challenges with the estimation of very late survival chances. The second assumption will ensure that the (resampled) estimators to be considered below have positive variances. The third assumption is required for reasons of identifiability and it is always achievable by artificially increasing \(\tau\) by a very small number.

As censoring is omnipresent in many medical studies, there often is no straightforward way to estimate \(\theta\). We assume that survival times are independently right-censored, that is, it is only possible to observe an event if it occurred before a so-called censoring time, say \(C_1\) and \(C_2\), respectively. These are allowed to be dependent whereas \((C_1, C_2)\) and \((T_1, T_2)\) are assumed to be independent. For the identifiability of \(\theta\), we assume that the square \([0, \tau] \times [0, \tau]\) is in the interior of the support of the joint distribution of \((C_1, C_2)\).

Finally, we assume that the actually observable data are of the form \((X_1, \delta_1, X_2, \delta_2)\), where \(X_j = \min(T_j, C_j, \tau)\) and \(\delta_j = 1(\min(T_j, \tau) \leq C_j), j = 1, 2\); here, \(1(\cdot)\) denotes the indicator function.

Note that the probability \(P(T_1 > \tau, T_2 > \tau)\) is implicitly included in the second part of the relative treatment effect; in order to estimate it and also the other involved probabilities in a unified way, we consider the case of \(X_j = \tau\) as uncensored, for technical reasons; see Section 2.2 for details. This does not mean that we actually assume the occurrence of the event at time \(\tau\). On a side note, Efron\(^{27}\) also regarded the largest observation as uncensored in order to achieve a so-called “self-consistency” property for the Kaplan-Meier estimator; see Section 7 therein.
After these preparations, we introduce the data modeled as \((X_{i1}, \delta_{i1}, X_{i2}, \delta_{i2}), i = 1, \ldots, n\), which are assumed to be independent and identically distributed. The main challenges in this article are the estimation of \(\theta\) from this censored paired time-to-event data set, and to draw reliable inferences about \(\theta\).

In the following, we will assume without loss of generality that the censoring times are continuously distributed prior to \(r\). In the case of discrete components in their distribution, ties can be broken by adding very small positive random numbers to them. These random numbers can be chosen small enough so that the order of all event times among the censoring times is not altered. Also, none of the statistical procedures considered below is affected by these small modifications of the censoring times.

### 2.2 Transformation of paired survival data into competing risks data, and estimation

Let us now describe our novel estimation approach for the relative treatment effect. For a facilitated estimation, we transform the paired survival data into a competing risks data set. The transformation is similar in spirit to the one used by Scheike et al.\(^\text{28}\) therein, a transformation of a paired competing risks data set into a univariate one facilitated the estimation of a concordance function. Our transformation works as follows: if, for a pair of right-censored event times,

1. the first entry is observed to fail before the second, we say: “an event of type 1 occurred”;
2. the second entry is observed to fail before the first, we say: “an event of type 2 occurred”;
3. both entries are observed to fail simultaneously, we say: “an event of type 3 occurred.”

All other cases are labeled “right-censored,” for example, if the observable event time of the first member of a pair is bigger than the observable censoring time of the second member. For each pair of event times, the competing risks event time is set to be the minimum of all four times: the minimum of both event times and both censoring times. We summarize the resulting competing risks data set as \((Z_i, \varepsilon_i) = (\min(\hat{T}_i, \hat{C}_i), \varepsilon_i \cdot 1\{\hat{T}_i \leq \hat{C}_i\}), i = 1, \ldots, n\). Here, for each \(i\), \(\hat{T}_i\) is the minimum of both event times, \(\hat{C}_i\) is the minimum of both censoring times, and \(\varepsilon_i\) denotes the event type. Note that the event type \(\varepsilon_i\) is unobservable in the case of a censoring. More technical details about the transformation are given in Section 7 in the Supplementary Material.

Denote by \(F_j(t) = P(\hat{T} \leq t, \varepsilon = j), t \in [0, r]\), the \(j\)th cumulative incidence function, \(j = 1, 2, 3\). A popular estimator for \(F_j\) is the Aalen-Johansen estimator,\(^\text{29}\) say \(\hat{F}_{j,n}\). The following proposition summarizes important implications of this artificial competing risks setting, and it introduces our proposed estimator \(\hat{\theta}_n\) of the relative treatment effect \(\theta\).

**Proposition 1.**

(a) (Representation) The relative treatment effect can be written as \(\theta = F_2(r) + \frac{1}{2}F_3(r)\).

(b) (Sufficiency) The above-described data reduction is sufficient for \(\theta\).

(c) (Efficiency) The following estimator is the nonparametric maximum likelihood estimator (NPMLE) of \(\theta\):

\[
\hat{\theta}_n = \hat{F}_{2,n}(r) + \frac{1}{2}\hat{F}_{3,n}(r).
\]  

A more specific structure of the estimator \(\hat{\theta}_n\) is given in the Supplementary Material. Many statistical properties of the Aalen-Johansen estimator \(\hat{F}_{j,n}\) are well-known, and they have direct implications for \(\hat{\theta}_n\), for example, consistency and asymptotic normality; cf. Andersen et al.\(^\text{30}\) Section IV.4. To summarize this, we denote convergences in probability and in distribution by \(p\rightarrow\) and \(\rightarrow\), respectively. The following theorem holds under the model specifications given in Remark 2.

**Theorem 1.** As \(n \rightarrow \infty\), we have \(\hat{\theta}_n \xrightarrow{p} \theta\) and \(\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \sigma_{\theta}^2)\) with asymptotic variance \(\sigma_{\theta}^2 = \sigma_2^2 + \sigma_3^2 + \frac{1}{2}\sigma_3^2 \in (0, \infty)\). Here, \(\sigma_2^2\) and \(\sigma_3^2\), respectively, denote the asymptotic variances of \(\sqrt{n}(\hat{F}_{2,n}(r) - F_j(r))\) and the asymptotic variances of \(\sqrt{n}(\hat{F}_{j,n}(r) - F_j(r))\), \(j = 1, 2, 3\).

A more detailed formula for the asymptotic variance \(\sigma_{\theta}^2\) is offered in the Supplementary Material. One possibility to estimate it is via the Greenwood-type (co)variance estimators of the Aalen-Johansen estimators.\(^\text{31,32}\) Using normal approximations for the statistics’ distributions, the results of Theorem 1 can be used to construct Wald-type tests and related confidence intervals. However, such inference procedures typically exhibit a suboptimal control of the type-I error.
rate and the confidence level, respectively, especially for small sample sizes. That is why we propose a resampling-based approach in Section 3 below.

2.3 | Discussion of related approaches in the literature

Let us review the present approach in the light of existing approaches and suggestions from the literature. Seigel and Podgor\(^6\) proposed to compare the counting processes for the competing risks of type 1 and 2 at time \(\tau\), say \(N_1(\tau) - N_2(\tau)\), that is, McNemar’s statistic. A statistical analysis of this difference would require taking the censoring rate into account. In contrast, the relative treatment effect estimator (2) relates to the difference of cumulative incidence functions through a one-to-one mapping: \(F_1(\tau) - F_2(\tau) = (1 - \theta) - \theta = 1 - 2\theta\). Similarly, \(1 - 2\hat{\theta}_n = \hat{F}_{1,n}(\tau) - \hat{F}_{2,n}(\tau)\). The latter may be written as \(\int_0^{\tau} (N_1 - N_2)(du)/\hat{G}_n(u-),\) where the denominator is the left-continuous version of the Kaplan-Meier estimator of the censoring survival function (in the competing risks data set). In that sense, our approach may be called an inverse-probability-of-censoring-weighting (IPCW)\(^33\) version of the suggestion by Seigel and Podgor.\(^6\)

The previous representation also illustrates the difference from the class of test statistics suggested by Dabrowska:\(^9\) she proposed \(T = \int K_n(s)(\hat{N}_1 - \hat{N}_2)(ds) + \int K_n(s)(\hat{N}_3 - \hat{N}_4)(ds)\). Here \(\hat{N}_1\) and \(\hat{N}_2\) are defined like \(\hat{N}_1, \hat{N}_2\), just based on the completely uncensored pairs, \(\hat{N}_3\) and \(\hat{N}_4\) correspond to the singly censored data points, and \(K_u\) and \(K_c\) are some scoring processes.

Let us also compare the present method to the Kaplan-Meier estimator-based approach by Dobler.\(^20\) Therein, a different variant of the relative treatment effect is analyzed, say \(N_1(\tau) - N_2(\tau)\), that is, McNemar’s statistic. A statistical analysis of this difference would require taking the censoring rate into account. In contrast, the method developed in the present paper is based on a direct comparison within each pair. In this sense, confounding is avoided or at least reduced. Thus, the novel competing risks-based approach will make better use of the available information about the dependence structure within the paired observations. It has been argued\(^34,36\) that it is challenging to draw causal conclusions for \(\hat{\theta}\) if not paired but only sample-specific measurements are available. In addition, Example 1 below illustrates that \(\hat{\theta}\) might exhibit some undesirable properties which do not occur for \(\theta\).

Finally, the parameter \(\hat{\theta}\) from (1) is related to the area under the ROC curve (AUC) which is sometimes presented similarly.\(^37\)

**Example 1.** We wish to illustrate some differences between the estimands \(\theta\) and \(\hat{\theta}\) in addition to the examples from the literature.\(^34,36\) In particular, we will point out a case with important implications for subpopulations. Let us suppose that the following data set is fully observable; a variant of the following examples has been kindly provided by Katharina Kramer (University of Augsburg). In that completely observable case without ties, the estimators simplify to the empirical fractions: \(\hat{\theta}_n\) is equivalent to the sign test statistic and \(\hat{\theta}_n\) is equivalent to the Mann-Whitney \(U\) test statistic.

| Pair | \(T_1\) | \(T_2\) | sign(\(T_1 - T_2\)) | Subgroup |
|------|--------|--------|---------------------|----------|
| 1    | 2      | 1      | +1                  | 1        |
| 2    | 4      | 3      | +1                  | 1        |
| 3    | 6      | 5      | +1                  | 2        |
| 4    | 8      | 7      | +1                  | 2        |

Here, we also suppose that additional information (a subgroup) is available, for example, males and females. For simplicity, let \(\tau = \infty\). In the example above, the above-mentioned estimators of \(\theta\) and \(\hat{\theta}\) yield the estimates \(\hat{\theta}_n = 1\) and \(\hat{\theta}_n = \frac{10}{16}\) respectively. These are the estimates for the whole population, that is, both subgroups combined. Within the subsamples, however, the estimates are \(\hat{\theta}_{j,n} = 1\) and \(\hat{\theta}_{j,n} = \frac{3}{4}, j = 1, 2\). Extending
this example reveals that \( \hat{\theta}_n \) can get arbitrarily close to \( \frac{1}{2} \) for the whole population\(^1\), that is, basically no Treatment 1 benefit, whereas \( \hat{\theta}_{1,n} = \frac{3}{4} > \frac{1}{2} \) in each subgroup consisting of exactly two pairs of event times. Note that a similar phenomenon cannot happen with \( \hat{\theta}_n \) which, in the fully observable case, is a convex combination of the subgroup-specific \( \hat{\theta}_{1,n} \) and \( \hat{\theta}_{2,n} \).

As a reviewer pointed out, the just-discussed data example relates to Simpson’s paradox because strong trends within each subgroup might be leveled out in the overall population. Another phenomenon that might complicate interpretations of \( \hat{\theta} \) in a \( k \)-sample setting (\( k > 2 \)) is the possible non-transitivity; one famous example is known as Efron’s dice\(^3\): it is possible that the relative treatment effect of sample \( j \) with respect to sample \( j + 1 \) is greater than 50\% for each \( j = 1, \ldots, k - 1 \), although the relative treatment effect of sample \( k \) with respect to sample 1 is also greater than 50\%. One solution to facilitating the interpretation in such a case is to compare each sample with the average of all sample groups.\(^39\) Note that this challenge might also occur with \( \theta \) in the case of matched triplets or bigger tuples.

It might also be useful to complement the estimand \( \theta \) with an average treatment effect, for example, differences of the restricted mean survival times \( E(\min(T_1, \tau) - \min(T_2, \tau)) \) to get additional insight in the effectiveness of a treatment with respect to another one.

3 \qquad INFERENCE: HYPOTHESIS TESTS AND CONFIDENCE INTERVALS

3.1 \qquad Overview

At the end of Section 2.2, we mentioned inference methods about the relative treatment effect \( \theta \in (0, 1) \). These can be based on the results of Theorem 1 in combination with a consistent variance estimator, \( \hat{\sigma}_n^2 \rightarrow \sigma_\theta^2 \) as \( n \rightarrow \infty \), and a normal approximation for the probability distribution of the estimator. In the following Section 3.2, we will improve the quality of the approximation by means of resampling methods. To this end, we will only look at (studentized) statistics in this subsection. In Section 3.3, we will discuss the role of transformations applied to the relative treatment effect. Concrete forms of hypothesis tests and confidence intervals will be given in Section 3.4. There, we will also point out why the confidence intervals and tests are coherent with each other, that is, a hypothetical value \( \theta_0 \) of the relative treatment effect is rejected by the test if and only if \( \theta_0 \) is not contained in a corresponding confidence interval. Choices of \( \theta_0 \) other than 0.5 could be relevant in practical applications, for example, when the treatment benefit should be sufficiently big to justify side effects from a drug or to justify high costs of a therapy. We will discuss finite exactness properties of the tests and confidence intervals under treatment exchangeability in Section 3.5.

3.2 \qquad Resampling-based inference

One well-known resampling option is a variant of the classical bootstrap,\(^40\)\(^41\) that is, draw \( n \) times independently with replacement from the competing risks data pairs \((Z_i, \epsilon_i), i = 1, \ldots, n\), and recompute \( \hat{\theta}_n \) based on the drawn bootstrap sample; call it \( \hat{\theta}_n^* \). By repeating this procedure a large number \( B \) of times, the collection of the normalized bootstrapped relative treatment effect estimators, say \( t_{n,b}^* = \sqrt{n}(\hat{\theta}_{n,b}^* - \hat{\theta}_n) / \hat{\sigma}_n^* \), \( b = 1, \ldots, B \), can be used to estimate the distribution of \( t_n = t_n(\theta) = \sqrt{n}(\hat{\theta}_n - \theta) / \hat{\sigma}_n \), for example, the \((1 - \alpha)\)-quantile which is necessary for a right-tailed test. Here, \( \hat{\sigma}_n^2 \) is an estimator of the variance of \( \hat{\theta}_n \) and \( \hat{\sigma}_{n,b}^2 \) are the bootstrap counterparts of \( \hat{\sigma}_n^2 \). Note that it is generally advised to use studentized statistics such as \( t_n(0.5) \) for tests.\(^42\)

Another, perhaps less well-known resampling option, is given by data re-randomization. The general procedure is similar to the bootstrap, except that data points are not drawn with replacement but instead some other random variation is introduced which is related to an algebraic group structure. One popular example is random permutation of the data points in a two independent samples setting which leads to permutation tests. Randomization tests can be shown to be finely exact if the re-randomization procedure reflects the data generation process. Recent works revisited the finite exactness of randomization tests.\(^43\)\(^44\) Similarly, the finite exactness of confidence intervals can be shown under randomization-invariance (up to the discreteness of the randomization distribution). In addition, one can argue the asymptotic exactness of randomization tests (with finite exactness in special cases) even if the data generation process does not exactly match the re-randomization method.\(^20\)
Now, we will apply such a randomization approach for inference about the relative treatment effect $\theta$. To be more precise, the randomization approach we propose is given by a random treatment re-allocation of both the outcomes within each matched data point. To motivate this, consider for a moment the strong null hypothesis $H^\text{exch}_n$ that both treatments are completely exchangeable in every respect. This implies for the relative treatment effect that $\theta = 0.5$. By treatment exchangeability, we technically mean that $(X_1, \delta_1, X_2, \delta_2)$ has the same distribution as $(X_2, \delta_2, X_1, \delta_1)$. For example, this is implied by the combination of the following four conditions: a symmetric copula dependence structure of the paired survival times, say, $C_{T_1,T_2}(u_1, u_2) = C_{T_2,T_1}(u_2, u_1)$, $u_1, u_2 \in [0,1]$, for the copula $C_{T_1,T_2}$ of $(T_1, T_2)$, and similarly for the copula of the pairs of censoring times $(C_1, C_2)$; equal marginal distributions of $T_1$ and $T_2$, and also equal marginal distributions of $C_1$ and $C_2$. Now, the benefit of the random treatment re-allocation is that it artificially creates a scenario of treatment exchangeability of which the collected data can be understood to be a mere realization—if treatment exchangeability indeed holds in reality. This gives rise to the exactness of randomization-based inference procedures under special circumstances of exchangeability; see Section 3.5 below.

In general, however, the inference methods should only be approximately exact for increasing sample sizes. For example, it is possible that the marginal survival functions of $T_1$ and $T_2$ are crossing (hence no exchangeability) in spite of a relative treatment effect that is potentially $\theta = 0.5$; see the simulation scenarios in Section 4. Yet, it is appreciated that they often still exhibit a good control of the statistical error rates in such cases of no finite exactness.\(^{\text{20}}\)

Note that our data re-randomization is different from the approach used in classical permutation tests for two independent samples, as the present data consist of paired data, and the randomization is done within each pair; but the general procedure for creating inference methods is the same.\(^{\text{45}}\)

Let us make the randomization approach more explicit. In terms of a fixed competing risks data set $(Z_i, \varepsilon_i), i = 1, \ldots, n$, a random re-labeling of both treatments means that the times $T_i$ remain unchanged but every occurrence of a type-1 and type-2 event is randomly re-labeled a type-1 or type-2 event, each with probability 50%. Denote the resulting randomized data set by $(Z_i, \varepsilon_i), i = 1, \ldots, n$, and the resulting randomized relative treatment effect by $\hat{\theta}_n$. In practice, to denote different realizations of the random re-labeling, while the original data are kept fixed, we will again use the index $b = 1, \ldots, B$ in the subscript to indicate the randomization iteration.

To justify the conditional convergence in distribution of the randomized relative treatment effect estimator, we introduce the notation $L$ for the distribution of a random variable and let $d$ denote a distance which induces a metric on the space of distributions on $\mathbb{R}$, for example, the Prokhorov distance.\(^{\text{46}}\) Let $\hat{\sigma}_{\theta,n}$ and $\tilde{\sigma}_{\theta,n}$ be suitable consistent estimators of the asymptotic variance of $\hat{\theta}_n$ and of the conditional variance of $\hat{\theta}_n$, respectively; see Section 8 in the Supplementary Material for this article and Section 4 for the practical implementation. Furthermore, we wish to point out that $P\cdot\lim_{n \to \infty} \tilde{\sigma}_{\theta,n}^2 = \sigma^2 \neq P\cdot\lim_{n \to \infty} \hat{\sigma}_{\theta,n}^2 = \hat{\sigma}^2 > 0$ in general, where $P\cdot\lim$ denotes the limit in probability. This underlines the necessity to studentize the (randomized) estimator. We will now state our main theorem for justifying the asymptotic correctness of the randomization approach, also when the treatment exchangeability does not hold. The theorem holds under the model specifications given in Remark 2.

**Theorem 2.** As $n \to \infty$,

$$d\left(L\left(\frac{\sqrt{n}(\hat{\theta}_n - 0.5)}{\tilde{\sigma}_{\theta,n}} \mid Z_1, \varepsilon_1, Z_2, \varepsilon_2, \ldots\right), L\left(\frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\hat{\sigma}_{\theta,n}}\right)\right) \to 0.$$

This theorem guarantees that the conditional distribution of $\tilde{t}_n = \sqrt{n}(\hat{\theta}_n - 0.5)/\tilde{\sigma}_{\theta,n}$ always approaches the distribution of $t_n = \sqrt{n}(\hat{\theta}_n - \theta)/\hat{\sigma}_{\theta,n}$ as $n \to \infty$, irrespective of the true value of $\theta$. This aspect is crucial for the development of confidence intervals for $\theta$ and tests for hypotheses, such as $H_{0,\theta_0} : \theta \leq \theta_0$, for some $\theta_0 \in (0, 1)$. Their actual confidence levels and type-I error probabilities will converge to the nominal $(1 - \alpha)$ and $\alpha \in (0, 1)$, respectively, as the sample size increases. Another consequence of Theorem 2 is that these tests are consistent under alternatives, that is, for each $\theta > \theta_0$, the power will tend to 1 with $n \to \infty$.

Finally, like for the bootstrap method explained above, the randomization method requires $B$ independent realizations of $\tilde{t}_n$, say, $\tilde{t}_{n,1}, \ldots, \tilde{t}_{n,B}$. Typically, $B$ should be a large number, for example, at least 1000. Then, quantiles of these realizations can be used for inferential purposes; alternatively, their empirical distribution, say $F_{\tilde{t},B,n}$, can be used to derive $P$-values for the randomized treatment-based tests: left, right, and two-tailed randomization-based $P$-values are respectively given...
by \( \hat{p}_1 = F_{I_{B,B,n}}(t_n(\theta_0)) \), \( \hat{p}_r = 1 - F_{I_{B,B,n}}(t_n(\theta_0)) \), and \( 2 \cdot \min(\hat{p}_1, \hat{p}_r) \). Bootstrap-based \( P \)-values can be derived similarly by replacing \( F_{I_{B,B,n}} \) with the empirical distribution of the realizations of the bootstrapped statistic \( t_n^{\phi,0} \). Asymptotic \( P \)-values result from replacing \( F_{I_{B,B,n}} \) with the cumulative distribution function of the standard normal distribution.

### 3.3 Transformation-based inference

Often, statistical properties of inference methods are improved by applying transformations, for example, variance-stabilizing transformations. In the present context, transformations could be used to ensure that confidence intervals for \( \theta \) are contained within the interval [0,1], or to further improve the type-I error control of a hypothesis test. To this end, let \( \phi : (0,1) \rightarrow \mathbb{R} \) be a continuously differentiable and monotonically increasing function, for example, \( \phi(\theta) = -\log(-\log(\theta)) \). A test for \( H_{0,\theta_0} : \theta \leq \theta_0 \) vs \( H_{A,\theta_0} : \theta > \theta_0 \) may be based on the test statistic \( t_n^{\phi,0}(\theta_0) = \sqrt{n}[\phi(\hat{\theta}_n) - \phi(\theta_0)]/\{(\phi'(\hat{\theta}_n)) \cdot \hat{\sigma}_{\theta,n}\} \); its randomization version is given by \( \tilde{t}_n^{\phi} = \sqrt{n}[\phi(\hat{\theta}_n) - \phi(0.5)]/\{(\phi'(\hat{\theta}_n)) \cdot \hat{\sigma}_{\theta,n}\} \). The denominators of the (randomization-based) test statistics are motivated by the \( \delta \)-method.

The \( \delta \)-method in combination with Slutzky’s theorem justifies the asymptotic correctness of the inference procedures derived from these (randomization-based) statistics. For a version of the \( \delta \)-method that applies to the randomization-based test statistic, we refer to Theorem 2 by Dobler. In the following Section 3.4, we will make the forms of the confidence intervals more explicit.

To describe the transformation and randomization-based \( P \)-values, let \( F_{I_{B,B,n}}^{\phi,0} \) be the empirical distribution of \( B \) independent realizations of \( t_n^{\phi,0} \). Then, left, right, and two-tailed \( P \)-values are respectively given by \( \tilde{p}_1^{\phi} = F_{I_{B,B,n}}^{\phi,0}(t_n^{\phi,0}(\theta_0)) \), \( \tilde{p}_r^{\phi} = 1 - F_{I_{B,B,n}}^{\phi,0}(t_n^{\phi,0}(\theta_0)) \), and \( 2 \cdot \min(\tilde{p}_1^{\phi}, \tilde{p}_r^{\phi}) \).

### 3.4 Coherence of tests and confidence intervals

We denote by \( q_{1-a,n}^{\phi,0} \) the unknown \( (1 - \alpha) \)-quantile of the unknown distribution of \( t_n^{\phi,0}(\theta_0) = \sqrt{n}[\phi(\hat{\theta}_n) - \phi(\theta_0)]/\{(\phi'(\hat{\theta}_n)) \cdot \hat{\sigma}_{\theta,n}\} \). Let us assume for simplicity that \( t_n^{\phi,0}(\theta_0) \) has a continuous distribution for each value of the true \( \theta \). Then, the test which rejects \( H_{0,\theta_0} : \theta \leq \theta_0 \) in favor of \( H_{A,\theta_0} : \theta > \theta_0 \) when \( t_n^{\phi,0}(\theta_0) > q_{1-a,n}^{\phi,0} \) has a type-I error probability equal to \( \alpha \) for each sample size \( n \). We can translate this into a statement for confidence intervals:

\[
1 - \alpha = P(t_n^{\phi}(\theta) \leq q_{1-a,n}^{\phi,0}) = P(-\phi(\theta) \leq -\phi(\hat{\theta}_n) + \phi'(\hat{\theta}_n) \cdot \hat{\sigma}_{\theta,n} \cdot q_{1-a,n}^{\phi,0}/\sqrt{n}) = P(\theta \geq \phi^{-1}(\phi(\hat{\theta}_n) - \phi'(\hat{\theta}_n) \cdot \hat{\sigma}_{\theta,n} \cdot q_{1-a,n}^{\phi,0}/\sqrt{n})).
\]

This reveals that the interval \( [\phi^{-1}(\phi(\hat{\theta}_n) - \phi'(\hat{\theta}_n) \cdot \hat{\sigma}_{\theta,n} \cdot q_{1-a,n}^{\phi,0}/\sqrt{n}), 1] \) covers \( \theta \) with probability \( 1 - \alpha \).

On the other hand, if we started with a confidence interval which exhibits the above structure, the display above would also reveal that the test obtained by rejecting all values of \( \theta_0 \) which are not contained in the confidence interval is a level \( \alpha \) test for \( H_{0,\theta_0} \) vs \( H_{A,\theta_0} \). Thus, such tests and confidence intervals yield results that are coherent with each other. This similarly holds for the corresponding two-sided versions and for lower confidence intervals and left-tailed tests. The just-described link between the inference methods is also known as the duality between tests and confidence intervals.

As explained in the previous sections, the exact quantiles \( q_{1-a,n}^{\phi,0} \) are unknown and have to be approximated. As a consequence, the equality on the left-hand side of (4) is an approximation at best, but the coherence between the tests and confidence intervals still holds. Note that incoherence may only arise if one uses the randomized version \( \tilde{c}_{1-a,n,B}^{\phi,0} \) of the hypothesis test.

### 3.5 Exactness of randomization-based tests and confidence intervals under exchangeability

Let us revisit (4) to understand the exactness of the randomization-based tests and confidence intervals under treatment group exchangeability. That is, \( q_{1-a,n}^{\phi,0} \) should be replaced by the randomization-based quantile, say \( \tilde{c}_{1-a,n,B}^{\phi,0} \) of \( B \).
independent realizations of $\tilde{t}_n^\phi$. Under treatment group exchangeability, $t_n^\phi(0.5)$ can be regarded to be one specific realization of $\tilde{t}_n^\phi$. In this sense, all realizations of $\tilde{t}_n^\phi$ are equally likely to be equal to $t_n^\phi(0.5)$. This is precisely the reason why the equality in (4) still holds for the randomization-based inference methods. As a consequence of (4), that is, the duality between tests and confidence intervals, both inference methods are exact for finite sample sizes $n$ whenever the treatment group exchangeability holds.

4 | SIMULATION STUDIES

4.1 | Section overview and simulation settings

The finite sample properties of the estimation and inference methods are analyzed with the help of simulation studies. Next to simulations for assessing the sizes of the proposed right-tailed tests under the null hypothesis $H_0 : \theta = 0.5$, we also ran power simulations and we simulated the coverage probabilities of confidence intervals for $\theta$. All simulations were conducted under R version 4.1.2. We first start with a description of the simulation settings in Section 4.1. In Section 4.2, we assess biases and root mean squared errors (RMSEs) of the relative treatment effect estimator as well as the sizes of the tests when $H_0 : \theta = 0.5$ is true. Section 4.3 deals with confidence intervals. In Section 4.4, we draw conclusions about the reliability of the inference methods for the relative treatment effect. Finally, in Section 4.5, we discuss the power of the tests when $H_0 : \theta > 0.5$ is true, also compared to other existing tests. The Supplementary Material contains additional simulation results.

In our simulation studies, we only included non-exchangeably distributed data in order to have a fair comparison between the non-randomized asymptotic, bootstrap, and randomization tests. That is, the randomization tests will not be exact level-\(\alpha\) tests. We considered the following combinations of simulation scenarios:

- \(n \in \{25, 50, 75, 100, 125, 150\}\) independent matched pairs, that is, \(2n\) (right-censored) event times in total;
- significance levels for the hypothesis tests: \(\alpha \in \{1\%, 5\%, 10\%\}\); only the results for \(\alpha = 5\%\) are included in the main body of this article, the others in Section 10.1 of the Supplementary Material;
- confidence level \(1 - \alpha = 95\%\) for confidence intervals for \(\theta\);
- copulas: Gumbel-Hougaard with parameter equal to 5; Clayton with parameter equal to $-0.6$;
- marginal distributions: \(\text{Exp}(2)\) vs \(\frac{1}{2}\text{Exp}(3)\)-\(\text{Exp}(\lambda)\)-mixture; \(\text{Gompertz}(0.6, b)\) vs \(\text{Exp}(3)\);
- censoring distributions (the same in both treatment groups): \(U(0, a)\) with \(a \in \{1.1, 1.6, 2.7\}\) for the first combination of marginal distributions, and \(a \in \{0.7, 1.175\}\) for the second combination;
- terminal time point \(\tau = 1\) for the first combination of marginal distributions, \(\tau = 0.6\) for the second combination;
- 5000 iterations of each test and confidence interval, each based on 2000 bootstrap and randomization iterations, respectively;
- 10 000 simulated data sets for each setting to assess the bias and the RMSE of $\hat{\theta}_n$.

The above-indicated rate parameters \(\lambda\) and \(b\) of the marginal distributions were found through numerous generations of large data sets and they were chosen such that \(\theta \approx 0.5\), that is, the null hypothesis is considered true. That is, under the Gumbel-Hougaard copula, \(b \approx 3.03\) and \(\lambda \approx 1.33\) and, under the Clayton copula, \(b \approx 3.16\) and \(\lambda \approx 1.30\). The parameters of the censoring distributions resulted in censoring rates of 38\% to 42\% (strong), 27\% to 34\% (medium), and 17\% to 27\% (light); these rates were found through simulations, where truncations at \(\tau\) were also considered censorings.

Let us explain the simulation procedure. First we generated the bivariate copula data. Next, the quantile functions of the marginal distributions are used to transform the copula data into bivariate data with the pre-selected dependence structure and marginal distributions. Finally, censoring is introduced by taking the minimum of the event and the simulated censoring times. For the purpose of estimating the relative treatment effect, the synthetic data are next transformed into a competing risks data set as described in Section 2.2. It should be pointed out that the cumulative incidence functions underlying the transformed competing risks data set neither have to be specified, nor do they play an important role, except for at time \(\tau\).
Table 1 contains the biases and RMSEs of the estimator $\hat{\theta}_n$ of the relative treatment effect on the scale of percentage points, that is, multiplied with 100. First, we notice that nearly all biases are negative. We generally see very small biases (ie, close to 0) under light censoring and, in most cases, still rather small biases under medium censoring. Under strong censoring, however, it increases (in absolute value), especially for small sample sizes. The combination of the Gumbel-Hougaard copula with exponential and exponential mixture marginals resulted in the largest biases (in absolute value). Under the medium and strong censoring regimes, the biases are more strongly pronounced when the dependence structure is given by the Gumbel-Hougaard copula. That is, under the strong positive correlation of the paired event times.

The previous two findings are also true in terms of the RMSEs. Especially for very small sample sizes $n \leq 50$ and under strong censoring, the RMSEs seem rather big. For $n = 150$, however, all but one RMSEs have decreased to values between 0.044 and 0.065. Below, we will evaluate the consequences in terms of the sizes of the hypothesis tests. Not surprisingly, all biases and RMSEs tend to decrease (in absolute value) as the sample size increases.

Table 2 contains the simulation results for the sizes of the right-tailed tests about the relative treatment effect $\theta$ when $H_0 : \theta = 0.5$ is true. They can be summarized as follows. Among the tests based on the untransformed relative treatment
effect ("lin." in the table), the asymptotic test tends to be a bit liberal, while the bootstrap test is somewhat conservative. This is more pronounced for the smaller sample sizes up to \( n = 100 \). The randomization test also exhibits a rather liberal behavior. For larger sample sizes (\( n \in \{125, 150\} \)), the sizes of all tests approach the 5% level quite accurately in most scenarios.

One notable peculiarity is the combination of strong censoring, the Gumbel-Hougaard copula, and, in particular, Gompertz vs exponential marginals: here, the bootstrap test stays rather conservative, and the randomization test is very anti-conservative. There is a slight improvement when the sample size increases. In view of the relatively large bias and RMSE as previously seen in Table 1, the anti-conservativeness is not entirely surprising. What surprises a bit, though, is that not the worst biases and RMSEs (Gumbel-Hougaard copula with exponential and exponential mixture marginals) resulted in the worst sizes of the tests (Gumbel-Hougaard copula with Gompertz and exponential marginals).

In this scenario, the asymptotic test is also anti-conservative but it performs better than the randomization test. For smaller sample sizes (\( n \in \{25, 50\} \)), similar observations about the simulation results can be made for the related medium censoring case. Figure 2 illustrates why the above-discussed scenario is rather challenging: many data points are converted into censorings on the competing risks scale; the censoring rate amounts to about 52%. At the same time, there is a very strong correlation between the survival times of a pair, and the marginal distributions are quite different. For smaller sample sizes, most of these characteristics are hardly visible. That is why this seems to be the most challenging simulation setting.

The log-log-transformation ("tra." in the table) seems to rectify the liberality of the asymptotic test, making it rather conservative, and it reduces the conservativeness of the bootstrap test in most scenarios. However, the transformation has nearly no effect on the type-I error control of the randomization test. All in all, the transformation-based bootstrap test seems to control the significance level most accurately among all considered right-tailed tests. In order to draw a more holistic conclusion in Section 4.4 below, we will also take results from left-tailed tests and two-sided problems into account.

### 4.3 Confidence intervals for \( \theta \neq 0.5 \)

In this section, we will assess the actual coverage rates of two-sided confidence intervals of nominal level \( 1 - \alpha = 95\% \) for the relative treatment effect \( \theta \in (0, 1) \). We created scenarios with \( \theta \neq 0.5 \) by multiplying 0.8 to the event times of the first treatment group; this also resulted in somewhat smaller censoring rates in the first group. The resulting relative treatment effects are smaller than 0.5: the true values of \( \theta \) are contained in \([0.289, 0.458]\). They were found through simulation of \( 2 \times 10^7 \) uncensored event times in each setting, in which case \( \hat{\theta}_n \) is an unbiased estimator of \( \theta \).

We would like to point out that the complementary probabilities the empirical coverage probabilities of the two-sided confidence intervals equal the sizes of the two-sided tests for \( H^\theta_{0, \theta_0} : \theta = \theta_0 \). Similarly, the empirical coverage probabilities of upper one-sided confidence intervals for \( \theta \) were implicitly simulated in Section 4.2. We refer to Section 3.4 for the duality between tests and confidence intervals.

Table 3 contains the simulated coverage probabilities (CPs) of the two-sided confidence intervals for \( \theta \). For most scenarios, the transformation seems to improve the CPs, that is, they get closer to 95%; one apparent exception is given by bootstrap-based confidence intervals under strong censoring. Let us abbreviate the CPs of the transformation-based confidence intervals with \( \Phi \)-CPs from now on. Under light censoring regimes, all \( \Phi \)-CPs were between 92.6% and 97.3%. Under medium and, in particular, strong censoring, the discrepancies from 95% increase, especially for the asymptotic and bootstrap-based confidence intervals in combination with small sample sizes. The randomization-based confidence intervals, however, exhibit CPs which are overall closest to 95% by far. This can be confirmed visually by means of Figure 3 where boxplots of the simulated coverage probabilities of all considered types of confidence intervals are shown across all simulation scenarios.

Hence, we conclude that the randomization-based confidence intervals are most reliable. Similarly to the test results from Section 4.2, there is no apparent difference in the coverage probabilities between the linear and transformed confidence intervals. Additional simulation results for confidence intervals for \( \theta = 0.5 \) are available in the Supplementary Material.
### Table 2

Simulated sizes of the right-tailed tests (in %) with nominal significance level $\alpha = 5\%$.

| Copula | Distribution | $n$ | Light censoring | Medium censoring | Strong censoring |
|--------|--------------|-----|-----------------|------------------|------------------|
|        |              |     | asy. | bs. | rand. | asy. | bs. | rand. | asy. | bs. | rand. | asy. | bs. | rand. | asy. | bs. | rand. | asy. | bs. | rand. |
|        |              | 25  | 6.5  | 4.2 | 3.8  | 5.9  | 5.4  | 5.3  | 6.8  | 4.4 | 4.3 | 6.4 | 6.0 | 5.9 | 5.5 | 3.8 | 3.6 | 4.5 | 6.3 | 6.3 |
|        |              | 50  | 5.1  | 3.8 | 3.7  | 4.6  | 4.6  | 4.6  | 6.8  | 5.0 | 5.1 | 6.1 | 6.4 | 6.3 | 4.7 | 4.0 | 3.6 | 3.9 | 6.8 | 6.7 |
|        |              | 75  | 4.9  | 3.7 | 3.8  | 4.5  | 4.5  | 4.5  | 5.7  | 4.4 | 4.7 | 5.3 | 5.1 | 5.1 | 4.2 | 3.7 | 3.2 | 3.7 | 7.0 | 7.0 |
|        |              | 100 | 4.8  | 3.7 | 3.9  | 4.4  | 4.4  | 4.4  | 5.8  | 4.3 | 4.8 | 5.4 | 4.9 | 4.9 | 4.2 | 3.7 | 3.5 | 3.8 | 6.8 | 6.8 |
|        |              | 125 | 4.9  | 4.2 | 4.4  | 4.8  | 4.7  | 4.7  | 5.8  | 4.7 | 5.0 | 5.5 | 5.3 | 5.3 | 4.2 | 3.7 | 3.6 | 3.7 | 7.1 | 7.1 |
|        |              | 150 | 5.3  | 4.4 | 4.6  | 5.0  | 4.9  | 4.9  | 5.1  | 4.5 | 4.6 | 4.9 | 4.6 | 4.6 | 4.4 | 3.9 | 3.5 | 3.7 | 7.2 | 7.1 |
|        |              | 25  | 7.3  | 4.5 | 1.9  | 4.4  | 6.0  | 6.2  | 7.9  | 5.0 | 2.7 | 5.0  | 8.3 | 8.5 | 8.0 | 4.9 | 3.2 | 4.3 | 14.9 | 14.7 |
|        |              | 50  | 5.7  | 3.9 | 3.4  | 4.6  | 4.9  | 4.9  | 5.7  | 3.9 | 3.1 | 4.1 | 6.1 | 6.1 | 5.9 | 4.2 | 2.5 | 4.0 | 13.0 | 13.1 |
|        |              | 75  | 4.9  | 4.1 | 4.1  | 4.6  | 4.7  | 4.7  | 5.4  | 4.1 | 3.6 | 4.3 | 5.5 | 5.5 | 4.5 | 3.4 | 2.1 | 3.0 | 11.7 | 11.7 |
|        |              | 100 | 4.7  | 3.8 | 4.0  | 4.6  | 4.4  | 4.4  | 4.9  | 3.7 | 3.6 | 4.1 | 5.0 | 5.0 | 4.4 | 3.6 | 2.8 | 3.2 | 11.0 | 11.1 |
|        |              | 125 | 5.2  | 4.3 | 4.6  | 5.0  | 5.0  | 5.0  | 5.4  | 4.5 | 4.7 | 5.1 | 5.3 | 5.4 | 4.3 | 3.2 | 2.6 | 3.2 | 10.4 | 10.5 |
|        |              | 150 | 5.3  | 4.5 | 4.8  | 5.2  | 5.2  | 5.2  | 5.3  | 4.3 | 4.5 | 4.9 | 5.2 | 5.2 | 3.7 | 3.1 | 2.8 | 3.2 | 9.7  | 9.8  |
|        |              | 25  | 5.7  | 3.5 | 2.8  | 5.1  | 4.5  | 4.5  | 6.3  | 3.4 | 3.0 | 5.5  | 4.7 | 4.7 | 5.8 | 3.3 | 3.1 | 5.3 | 4.8 | 4.7 |
|        |              | 50  | 5.4  | 3.8 | 3.7  | 4.7  | 4.6  | 4.6  | 5.6  | 4.1 | 3.9 | 4.8 | 5.0 | 5.1 | 5.1 | 3.8 | 3.5 | 4.3 | 5.0 | 5.1 |
|        |              | 75  | 5.3  | 4.2 | 4.2  | 4.7  | 5.0  | 5.0  | 5.3  | 4.3 | 4.4 | 4.9 | 5.4 | 5.4 | 5.3 | 4.2 | 4.2 | 4.7 | 5.4 | 5.3 |
|        |              | 100 | 5.7  | 4.7 | 4.7  | 5.2  | 5.6  | 5.6  | 5.1  | 4.4 | 4.4 | 4.8 | 5.2 | 5.2 | 5.3 | 4.3 | 4.1 | 4.6 | 5.9 | 5.9 |
|        |              | 125 | 5.6  | 4.8 | 5.0  | 5.4  | 5.6  | 5.6  | 4.5  | 3.7 | 3.8 | 4.1 | 4.6 | 4.6 | 4.7 | 3.9 | 3.8 | 4.2 | 5.0 | 5.1 |
|        |              | 150 | 4.8  | 4.2 | 4.3  | 4.6  | 4.8  | 4.8  | 5.1  | 4.4 | 4.4 | 4.7 | 5.3 | 5.3 | 3.9 | 3.4 | 3.2 | 3.4 | 4.7 | 4.7 |
|        |              | 25  | 6.6  | 3.9 | 2.7  | 5.2  | 5.0  | 5.1  | 6.8  | 4.0 | 3.1 | 5.8  | 5.1 | 5.1 | 6.8 | 3.9 | 3.0 | 6.1 | 5.6 | 5.6 |
|        |              | 50  | 6.0  | 4.6 | 4.3  | 5.1  | 5.2  | 5.2  | 5.9  | 4.2 | 3.6 | 4.9 | 5.2 | 5.2 | 6.3 | 4.4 | 4.0 | 5.0 | 5.7 | 5.7 |
|        |              | 75  | 5.4  | 4.4 | 4.2  | 4.8  | 4.8  | 4.8  | 5.6  | 4.2 | 4.0 | 4.7 | 5.0 | 5.0 | 5.8 | 4.4 | 3.7 | 4.5 | 6.0 | 6.0 |
|        |              | 100 | 5.7  | 4.6 | 4.7  | 5.2  | 5.4  | 5.4  | 5.7  | 4.7 | 4.5 | 5.0 | 5.5 | 5.5 | 5.6 | 4.7 | 4.5 | 5.0 | 5.8 | 5.8 |
|        |              | 125 | 5.5  | 4.8 | 4.9  | 5.2  | 5.3  | 5.3  | 5.7  | 4.6 | 4.6 | 5.0 | 5.3 | 5.3 | 5.6 | 4.7 | 4.2 | 4.7 | 5.8 | 5.8 |
|        |              | 150 | 5.4  | 4.9 | 5.0  | 5.3  | 5.3 | 5.3  | 5.1  | 4.3 | 4.4 | 4.6 | 4.9 | 4.9 | 5.5 | 4.5 | 4.4 | 4.7 | 5.5 | 5.5 |

Abbreviations: bs., bootstrap; Copula: GH, Gumbel-Hougaard; critical values: asy., asymptotic normal; rand., randomization; tests: lin., linear; tra., log-log-transformed.
| Copula  | Distribution              | n  | Light censoring | Medium censoring | Strong censoring |
|---------|---------------------------|----|-----------------|------------------|------------------|
|         |                           |    | asy. lin. tra.  | asy. lin. tra.   | asy. lin. tra.   |
|         |                           |    |                 |                  |                  |
|         |                           |    |                 |                  |                  |
| GH      | Exp vs Exp mix            | 25 | 91.8 95.1       | 98.0 97.3        | 88.2 93.2        |
|         |                           | 50 | 93.1 94.7       | 96.5 93.3        | 90.3 92.9        |
|         |                           | 75 | 93.8 94.8       | 96.3 94.5        | 91.2 92.6        |
|         |                           | 100| 93.6 94.3       | 95.5 94.7        | 91.2 92.5        |
|         |                           | 125| 93.9 94.3       | 95.4 94.6        | 92.1 93.1        |
|         |                           | 150| 94.2 94.7       | 95.4 94.8        | 92.5 93.3        |
|         |                           | 25 | 92.2 95.1       | 98.2 96.9        | 92.2 96.0        |
|         |                           | 50 | 93.8 95.2       | 96.5 94.2        | 93.7 95.2        |
|         |                           | 75 | 93.9 94.8       | 95.8 94.0        | 94.3 95.5        |
|         |                           | 100| 94.4 95.1       | 95.8 94.6        | 94.1 95.0        |
|         |                           | 125| 94.7 95.3       | 95.8 94.9        | 94.4 95.0        |
|         |                           | 150| 94.4 94.8       | 95.5 94.6        | 94.7 95.3        |
| GH      | Gompertz vs Exp           | 25 | 92.1 95.1       | 98.2 96.9        | 92.2 96.0        |
|         |                           | 50 | 93.8 95.2       | 96.5 94.2        | 93.7 95.2        |
|         |                           | 75 | 93.9 94.8       | 95.8 94.0        | 94.3 95.5        |
|         |                           | 100| 94.4 95.1       | 95.8 94.6        | 94.1 95.0        |
|         |                           | 125| 94.7 95.3       | 95.8 94.9        | 94.4 95.0        |
| GH      | Exp vs Exp mix            | 25 | 91.2 93.8       | 97.8 92.6        | 91.3 93.3        |
|         |                           | 50 | 93.3 94.6       | 96.4 94.3        | 93.0 94.1        |
|         |                           | 75 | 94.1 94.9       | 96.0 94.7        | 92.7 93.5        |
|         |                           | 100| 94.0 94.5       | 95.6 94.7        | 93.7 94.1        |
|         |                           | 125| 94.1 94.3       | 95.4 94.6        | 93.6 94.0        |
|         |                           | 150| 94.5 94.9       | 95.6 95.0        | 93.7 94.2        |
| GH      | Gompertz vs Exp           | 25 | 92.1 94.2       | 98.1 93.3        | 91.5 93.8        |
|         |                           | 50 | 93.0 94.2       | 96.3 94.4        | 93.0 94.5        |
|         |                           | 75 | 94.1 94.9       | 96.0 94.7        | 92.9 93.5        |
|         |                           | 100| 94.2 94.6       | 95.7 94.8        | 93.7 94.2        |
|         |                           | 125| 94.4 94.8       | 95.7 94.8        | 94.5 95.0        |
|         |                           | 150| 94.3 94.7       | 95.7 95.1        | 94.3 94.9        |

Abbreviations: bs., bootstrap; Copula: GH, Gumbel-Hougaard; critical values: asy., asymptotic normal; rand., randomization; tests: lin., linear; tra., log-log-transformed.
FIGURE 2 Scatterplots of $n = 200$ simulated data points according to a Gumbel-Hougaard copula and, respectively, Gumbel and exponentially distributed marginals. The red line is the diagonal $y = x$. The dashed lines illustrate the upper end of the censoring support $(0.7)$ and $r = 0.6$. Upper panels: Complete data set (no censoring); lower panels: censored data set. Left panels: raw data; censorings in both coordinates are denoted by “+”; censorings only in the horizontal coordinate are denoted “X”; censorings only in the vertical coordinate are denoted “Y”; completely uncensored data points are denoted by red circles. Right panels: competing risks data after transformation. The event time is the minimum of both coordinates. The symbols “1,” “2,” and “C” represent whether the data point corresponds to an observed event of type 1, type 2, or to a censoring. Type 3 events are also present but not depicted.

4.4 Summary of the simulation studies

Let us once again compare all inference methods: asymptotic, bootstrap-, and randomization-based hypothesis tests and confidence intervals. In general, we will interpret the coverage probabilities of the confidence intervals similarly to the type-I error rates of the tests, due to the duality between tests and confidence intervals; compare Section 3.4. Also, we found the transformation $\phi$ to be generally beneficial, so we will only focus on the transformation-based results.

In Section 4.2, we considered right-tailed tests. There, we found the asymptotic test to be conservative, the randomization test to be liberal and the bootstrap test to be most accurate. In Section 4.3, for the two-sided problem, we found the randomization-based confidence interval for $\theta \neq 0.5$ to be quite accurate, even under under strong censoring regimes. In
contrast, the asymptotic and bootstrap-based confidence intervals turned out to be extremely liberal in some of the settings. Similar observations can be made for the confidence intervals for $\theta = 0.5$ given in Table 7 in the Supplementary Material.

In fact, the (extreme) liberality of the asymptotic and the bootstrap-based confidence intervals is a consequence of the (extreme) liberality of the corresponding left-tailed tests. In contrast, also the left-tailed randomization tests appeared to be very accurate, albeit slightly conservative (results not shown).

All in all and apart from the right-tailed tests under settings with the Gumbel-Hougaard copula and strong censoring, the one- and two-sided randomization-based inference methods were most reliable across a majority of simulation scenarios. In the special case of strong censoring combined with a strong positive correlation between the paired event times though, one should be aware that the randomization-based methods are potentially somewhat liberal. However, a comparison of the corresponding simulation results for all one- and two-sided problems revealed that the randomization-based methods are still the most reliable ones. Because of the additional finite exactness property of the randomization-based methods under treatment exchangeability, our recommendation is to prefer resampling via randomization over the bootstrap. This should be combined with the transformation $\phi$ to ensure that the confidence intervals are contained in $[0, 1]$.

We will investigate the power of the recommended randomization test in the subsequent subsection.

4.5 Power simulations

Next to the simulations of the sizes of the test under $H_0 : \theta = 0.5$ and the coverage probabilities of the related confidence intervals for $\theta$, we have also conducted a simulation study to assess the power of the one-tailed version of the developed tests. Since the results for the transformed and untransformed test statistics are very much alike, we have solely focused on the latter. We considered two competitor tests, also in their one-tailed versions: the paired Prentice-Wilcoxon test\(^7\) which was found to be very powerful in the comparative simulation study by Woolson and O’Gorman\(^15\); the stratified log-rank test\(^11\) which was more closely analyzed under correlated frailty models. In order to ensure a fair comparison, the randomization versions of all tests were used, such that all of them control the significance level for finite sample sizes under exchangeability. In this subsection, we chose the significance level $\alpha = 5\%$.

We considered the same two copulas as in the previous subsection, the sample sizes $n \in \{25, 50, 100\}$, 1000 test replications, 1000 randomization iterations, and the following three marginal distribution scenarios:

1. Mixture of the Exp(2)-exponential and the $U(0, 2)$-uniform distribution against the Exp(2)-exponential distribution; the censoring times were independently $U(0, 2.5)$-distributed; $\tau = 1.9$. This scenario departs from the sharp null hypothesis of exchangeability into alternatives with crossing hazard rates at late time points close to $\tau$ as the mixing parameter puts more and more weight on the uniform distribution.
2. Mixture of the Exp(2)-exponential and the Gompertz distribution with shape parameter 0.1 and rate parameter 2 against the Exp(2)-exponential distribution; the censoring times were independently $U(0, 2.5)$-distributed; $r = 1.8$. This scenario departs from the sharp null hypothesis of exchangeability into alternatives with crossing hazard rates at central time points as the mixing parameter puts more and more weight on the Gompertz distribution.

3. The Exp($2/k$)-exponential distribution against the Exp(2)-exponential distribution; the censoring times were independently $U(0, 2)$-distributed; $r = 1.3$. This scenario departs from the sharp null hypothesis of exchangeability into alternatives with parallel hazard rates as the scale parameter $k$ increases from 1 to 2.

For now, we only focus on the results for Scenario 1 graphically presented in Figure 4; the results for Scenarios 2 and 3 are presented in Section 10.2 of the Supplementary Material are generally similar. We can see from Figure 4 by comparing each combination of the left and right panels that the copula that connects the lifetimes seems to have little influence on the performance of the tests. Not surprisingly, the powers of all tests increase when the sample size increases (top to bottom in the figure) and when we depart from the null hypothesis (from left to right within each panel). The paired Prentice-Wilcoxon test\textsuperscript{7} is always the most powerful one. In most cases, the proposed test has the next higher power but its performance is generally very similar to that of the stratified log-rank test.\textsuperscript{11}

Multiple comments are in order. A first reason for the relatively high power of the log-rank test and the paired Prentice-Wilcoxon test is that the generated lifetimes were not truncated at $r$, whereas smaller values of $r$ result in relative treatment effects $\theta$ closer to 0.5, that is, closer to the null hypothesis.

Second, in this simulation study, we could confirm the rather high power of the paired Prentice-Wilcoxon test that has earlier been reported in the literature.\textsuperscript{15} This is also not surprising because it is based on an efficient score approach. Next, the power of the stratified log-rank test could possibly be greater if the optimal combination with the unstratified log-rank test was used.\textsuperscript{11} However, an implementation of the combination would be beyond the scope of the present paper. In addition, it should not be forgotten that the considered competitor tests were proposed for the sharp null hypothesis of equal survival distributions, $H_0^\text{exch.} : S_1 \equiv S_2$ (vs the inequality or some strict ordering); it is only natural that such tests potentially have a greater power than the proposed test which was designed for the weak null hypothesis $H_0 : \theta = 0.5$, which is implied by $H_0^\text{exch.} : S_1 \equiv S_2$. The newly proposed test is thus more generally applicable. As a consequence, comparing these tests has a flavor of comparing a goodness-of-fit test for a simple hypothesis with one for a composite hypothesis. Nevertheless, it is only natural that the interested reader wishes to know the discrepancies of power in a situation where both tests are applicable and thus comparable. That is the sole reason for reporting the results of the power simulation for the competitor tests in the present paper.

Finally, to put all tests in a perspective, we wish to reiterate that treatment exchangeability ($H_0^\text{exch.}$) implies the equality of both marginal survival functions ($H_0^\text{exch.}$) which in turn implies a relative treatment effect of $\theta$ other than 0.5. In the light of all these points and the acceptable power of the proposed test procedure, our novel test is a competitor with many benefits.

5 | DATA EXAMPLE

We illustrate our methodology by re-analyzing a well-known benchmark data set which has been published in the R package survival.\textsuperscript{46} The data set diabetic, in detail described and analyzed by the Diabetic Retinopathy Study Research Group\textsuperscript{49} and further investigated by Huster et al.,\textsuperscript{4} contains 394 observations from a trial including 197 patients with “high-risk” diabetic retinopathy, a complication associated with diabetes mellitus that frequently leads to blindness. In this trial each patient acts as his or her own control: one eye was randomized to a laser photocoagulation, while the other eye received no treatment.

Apart from suffering from diabetic retinopathy, the inclusion criterion of the trial was a visual acuity of at least 20/100 in both eyes. The aim of the study was to investigate the effect of the laser treatment on delaying the onset of blindness, defined as “severe visual loss,”\textsuperscript{50} that is, a visual acuity of less than 5/200 at two or more consecutively completed follow-up visits. These visits were scheduled at four months intervals, with a maximum follow-up time of 72 months. Thus, the survival times are given by the time (in months) from the beginning of the treatment until the occurrence of blindness. Censoring was caused by death, dropout, or the end of the study. Consequently, we consider the censored paired survival outcomes and their relative treatment effect.
The data set consists of two subgroups, defined by the type of diabetes, that is, patients with juvenile onset diabetes (diagnosis before an age of 20, i.e., 114 patients) and adult onset diabetes (83 patients). In these subgroups, respectively 78.9% and 85.1% of the patients had at least one censoring. Further, there are several other covariates, for example, the laser type and a risk score.

For a better overview of the data, Figure 5A visualizes the time until blindness for the total sample, regardless of the age at diagnosis. From the pattern of the observations displayed in the figure, one can see the tendency that the laser treatment generally seems to delay blindness. For a separate investigation of the juvenile and adult sample, respectively, Figure 5B displays the estimated Kaplan-Meier curves for each eye, that is, the one receive a laser treatment and the other one acting as the control. For both subsamples, we observe a visible difference between the Kaplan-Meier curves
FIGURE 5  Graphical summaries of the data set. (A) Times until blindness (subject to censoring) for the diabetic retinopathy data: the times related to the treated eyes are given in the horizontal dimension, the corresponding times of the control in the vertical dimension. The values on the diagonal (denoted by “+”) correspond to patients where no blindness occurred throughout the observational period. Red circles indicate blindness of both eyes, values below the diagonal (denoted by “Y”) indicate blindness of the control eye only, whereas values above the diagonal (denoted by “X”) indicate blindness of the treated eye only. (B) Kaplan-Meier-curves for the two eyes (treated and control) for the diabetic retinopathy data, fitted separately for the juvenile and the adult sample.

which supports our conjecture that the laser treatment seems to delay the onset of blindness compared to the control; this becomes even more visible in the adult sample.

In order to confirm our visual findings, we estimated the relative treatment effect of the laser photocoagulation and the corresponding confidence intervals, as well as performed the corresponding two-sided hypothesis tests in order to assess the effectiveness of this therapy. We fixed the maximum follow-up time as \( \tau = 60 \) months (indicated by the dashed box in Figure 5A) and performed the analysis for both samples, that is, juvenile and adult onset diabetes, separately. For the juvenile sample, we obtained a relative treatment effect of \( \hat{\theta}_J = 0.598 \) (estimated standard deviation: \( \hat{\sigma}_{\theta,J} = 0.041 \)); for the adult sample, we had \( \hat{\theta}_A = 0.731 \) (\( \hat{\sigma}_{\theta,A} = 0.038 \)). The corresponding 95\%-confidence intervals and the results of the (one-sided) hypothesis test for the different approaches described in Section 3 are summarized in Table 4. The transformation used for the analysis is given by \( \phi(\theta) = -\log(-\log(\theta)) \) and all results based on bootstrap and randomization were achieved by using \( B = 2000 \) resampling repetitions. For each subsample, all confidence intervals and also all P-values are very similar. In general, the effect of an additional transformation of the test statistic is rather small and yields very similar confidence intervals and test results. As already indicated by Figure 5B, there is a notable difference of treatment and control eye, which is even larger in the adult sample. This is confirmed by P-values below .001 for the adult sample for all tests under consideration. For the juvenile sample, the P-values lie between .008 and .012. Hence, we conclude a significant treatment effect for both samples at the significance level \( \alpha = 5\% \).

Our results are similar to the findings from Oakes and Feng,\(^{11} \) who investigated the same data set regarding the treatment effect. In their paper, they proposed three different test approaches and concluded that, for the example at hand, all resulting treatment effects were significant. Indeed, the right-tailed randomization-based tests by O’Brien and Fleming and Oakes and Feng as considered in Section 4.5 above resulted in P-values below .001 (both tests for the adult subgroup) and .017 and .016, respectively, for the juvenile subgroup. This is in line with the P-values found from the randomization-based test proposed in the present paper. However, in contrast to our method, the two just-mentioned competitors do not automatically also provide confidence intervals for a parameter that is easy to interpret.
### TABLE 4  Comparison of 95%-confidence intervals and right-tailed P-values for the diabetic retinopathy data.

| Method | Juvenile sample | Adult sample |
|--------|-----------------|--------------|
|        | 95%-confidence interval | P-value | 95%-confidence interval | P-value |
| asy. lin. | [0.518, 0.678] | .008 | [0.656, 0.806] | <.001 |
| tra. | [0.514, 0.672] | .122 | [0.647, 0.798] | <.001 |
| bs. lin. | [0.518, 0.683] | .011 | [0.656, 0.805] | <.001 |
| tra. | [0.520, 0.681] | .009 | [0.658, 0.803] | <.001 |
| rand. lin. | [0.517, 0.684] | .010 | [0.655, 0.807] | <.001 |
| tra. | [0.517, 0.680] | .011 | [0.652, 0.800] | <.001 |

Abbreviations: asy., asymptotic normal; bs., bootstrap; lin., linear; rand., randomization; tra., log-log-transformed.

## 6  DISCUSSION

In this article, we developed a new estimand in the context of paired, right-censored survival data, the so-called relative treatment effect, to compare the effectiveness of two treatments. Such data occur for instance in matched pairs studies. Our focus was on (interval) estimation and hypothesis tests for that parameter. We proved the asymptotic exactness of the confidence intervals and tests, that is, for increasing sample sizes, by means of a central limit theorem for the relative treatment effect estimator $\hat{\theta}_n$ and a similar result for its randomization version $\tilde{\theta}_n$. By means of simulation studies, we verified that the statistical reliability of the randomization-based inference methods in terms of type-I error rates and coverage probabilities is generally satisfactory in small to medium-size samples.

The relative treatment effect $\theta$ quantifies the stochastic ordering of treatment outcomes but not how much bigger one survival times is than the other. In this sense, it is a global measure for the superiority of the first treatment, although other measures, for example, about the actual size of the differences, might also be of major importance in some applications.

There are multiple possibilities for other extensions of the present approach. One open question is how to incorporate additional patients that could not be matched with others or if multiple patients of one treatment group could be matched with just one patient of the other group. The latter problem could potentially be approached by means of an appropriate re-weighting of the within-pair comparisons. However, there is the risk that the interpretability of the relative treatment effect could be lost.

Another way to extend the present approach is the incorporation of additional covariates. The inclusion of covariates in the analysis could be used to tackle the classification problem of who specifically should receive which treatment. Due to the favorable competing risks approach, such an extension could be achieved rather straightforwardly, for example, by means of cause-specific hazard models or subdistribution hazard models. Another approach could be to involve relative treatment effects in a machine learning algorithm. The use of additional covariates could also be used for causal inferences about $\theta$ for a target population for which the study population might not be entirely representative: because the within-pair comparisons eliminate or at least reduce confounding, again a re-weighting approach based on the covariates could be used to transfer conclusions for the study population to the actual target population.

Next, we note that the hypothesis test presented in this article investigates the significance of the treatment effect. However, there might also occur situations where one is rather interested in investigating whether the deviation of the treatment effect of 0.5 is not larger than pre-specified values $\epsilon_1$ and $\epsilon_2$, respectively. In other words, this requires an equivalence test for $H_0^{eq}: \theta \not\in (0.5 - \epsilon_1, 0.5 + \epsilon_2)$ against $H_1^{eq}: \theta \in (0.5 - \epsilon_1, 0.5 + \epsilon_2)$; see, for example, the monograph by Wellek. Such an approach could provide a very flexible framework for statistical inference, address numerous other research questions, and consequently provide a useful addition to the test proposed in this article. We leave the development of such a procedure for future research.

Finally, in view of the real data analysis presented in this article, we would like to conclude with a word of caution. A reviewer discussed the modeling of death as one cause of censoring. And we agree with him or her that, instead, one could regard this as a competing risk. In this case, one may define an endpoint-specific relative treatment effect for each possible combination of endpoints within each matched pair, for example, (death, death), (blindness, death) and so forth. This is a good point for future research. To come back to the role of death as censoring, mortality and blindness indeed seem positively correlated. As for the mortality in the diabetic retinopathy study, the 1-yr and 5-yr cumulative mortality
rates were 3.0 and 22.6, respectively. As a consequence, it seems that one should not ignore the potential problems induced by dependent censoring. For the illustration of our novel methodology, however, we found the analysis of the diabetic retinopathy data set appropriate: it offers a useful illustration of a matched pairs problem that arises naturally; it is freely available and the results are reproducible; also, due to the availability of additional covariates, it offers many possibilities for methodological extensions. In addition, our analysis complements the findings of an earlier analysis of these data which appeared in Statistics in Medicine.

Finally, motivated by the just-initiated discussion about dependent censoring in the present context, we would like to point out another interesting point for future research: an extension of the present method to the case of dependent censoring. In a recent work, Emura et al developed statistical tests for a multi-sample version of $\theta$ in the context of general factorial designs and dependent right-censoring. Hence, a combination of their approaches and the present paper could turn out to be fruitful.

ACKNOWLEDGEMENTS
The authors wish to thank Arthur Allignol and Mark Clements for providing an update of the etm R package, Marialuisa Restaino, Nan van Geloven, Liesbeth de Wreede, Hein Putter, Giuliana Cortese, Thomas Scheike, Katharina Kramer, Merle Munko, and Paul Blanche for useful discussions. In addition, the comments, questions, and suggestions of an Associate Editor and three reviewers led to a substantial improvement of the quality of this article.

CONFLICT OF INTEREST STATEMENT
The authors declare that they have no conflict of interest.

DATA AVAILABILITY STATEMENT
The R code used for the simulation studies, for generating synthetical data sets, and for the real data application as well as a more detailed derivation of an equivalent variance (estimator) are available at https://github.com/dennis-dobler/relative_treatment_effect_paired_survival.

ENDNOTES
*A complication affecting the eyes which is associated with diabetes mellitus.

†When the $i$th pair of the data set contains the event times $(2i, 2i - 1)$, $i = 1, \ldots, n$, then $\tilde{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{n(n+1)/2} = \frac{n+1}{n} \cdot \frac{1}{2} \rightarrow \frac{1}{2}$ as $n \rightarrow \infty$.

§By a randomized version of a test, we mean a test which rejects the null hypothesis with a probability $\pi \in (0, 1)$ if the test statistic is equal to the critical value. This is not to be confused with the randomized test which results from the described data re-randomization procedure to compute the critical or $P$-value.

At least, up to the discreteness of the randomization distribution. In any case, $1 - \alpha \geq P_{n}^{\phi}(\theta) \leq c_{1-\alpha,n,B}$ can always be guaranteed under exchangeability. Using the randomized version of the randomization-based tests will result in the exact level of $\alpha$ under exchangeability.
12. Pacock SJ, Ariti CA, Collier TJ, Wang D. The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. *Eur Heart J.* 2012;33(2):176-182.

13. Holt JD, Prentice RL. Survival analyses in twin studies and matched pair experiments. *Biometrika.* 1974;61(1):17-30.

14. Jeong GJ. A covariate-adjusted logrank test for paired survival data. *Commun Stat Appl Methods.* 2002;9(2):533-542.

15. Woolson RF, O’Gorman PW. A comparison of several tests for censored paired data. *Stat Med.* 1992;11(2):193-208.

16. Le-Rademacher J, Brazauskas R. Inference for paired survival data. *Handbook of Survival Analysis.* 1st ed. New York: Chapman and Hall/CRC Press; 2013:615-632.

17. Brunner E, Vanhemelryck M, Mütze T. Win odds: an adaptation of the win ratio to include ties. *Stat Med.* 2021;40(14):3367-3384.

18. Munko M, Ditzhaus M, Dobler D, Genuneit J. RMST-based multiple contrast tests in general factorial designs. *Stat Med.* 2024;1-18. doi: 10.1002/sim.10017

19. European Medicines Agency. ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials; 2017.

20. Dobler D. Randomized empirical processes by algebraic groups, and tests for weak null hypotheses. *J Am Stat Assoc.* 2002;97(458):224-237.

21. Pruitt RS. Small sample comparison of six bivariate survival curve estimators. *J Stat Comput Simul.* 1993;45(3-4):147-167.

22. Duijtselaere J, Restaino M, Wang H. A class of nonparametric bivariate survival function estimators for randomly censored and truncated data. *J Nonparametr Stat.* 2016;28(4):736-751.

23. Brunner E, Munzel U. The nonparametric Behrens-Fisher problem: asymptotic theory and a small-sample approximation. *Biometrika.* 2000;42(1):17-25.

24. Dobler D. Randomized empirical processes by algebraic groups, and tests for weak null hypotheses. *Bernoulli.* 2023;29(2):1109-1136.

25. Munzel U, Brunner E. An exact paired rank test. *Biometrika.* 2002;44(5):584-593.

26. Konietzchke F, Pauly M. A studentized permutation test for the non-parametric Behrens-Fisher problem in paired data. *Electron J Stat.* 2012;6:1358-1372.

27. Efron B. The two sample problem with censored data. In: Le Cam LM, Neyman J, eds. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability.* Berkeley, CA: University of California Press; 1967:831-853.

28. Scheike TH, Holst KK, Hjelmborg JB. Estimating twin concordance for bivariate competing risks twin data. *Stat Med.* 2014;33(3):1193-1204.

29. Aalen OO, Johansen S. An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scand J Stat.* 1978;5(3):141-150.

30. Andersen PK, Borgan O, Gill RD, Keiding N. *Statistical Models Based on Counting Processes.* New York: Springer; 1993.

31. Allignol A, Schumacher M, Beyersmann J. A note on variance estimation of the Aalen–Johansen estimator of the cumulative incidence function in competing risks, with a view towards left-truncated data. *Biometrics.* 2012;68:526-534.

32. Robins JM, Rotnitzky A. Recovery of information and adjustment for dependent censoring using surrogate markers. In: Jewell N, Dietz K, Farewell V, eds. *AIDS Epidemiology—Methodological Issues.* Boston, MA: Birkhäuser; 1992:297-331.

33. Fay MP, Brittain EH, Shih JH, Follmann DA, Gabriel EE. Causal estimands and confidence intervals associated with Wilcoxon-Mann-Whitney tests in randomized experiments. *Stat Med.* 2018;37(20):2923-2937.

34. Hartman N, Kim S, He K, Kalbfleisch JD. Pitfalls of the concordance index for survival outcomes. *Stat Med.* 2023;42(13):2179-2190.

35. Greenland S, Follmann DA, Shih JH, Gabriel EE. Causal claims about the D-value. *Stat Probab Lett.* 2021;174:109274-109283.

36. Greenland S, Follmann DA, Shih JH, Gabriel EE. Causal claims about the D-value. *Stat Probab Lett.* 2021;174:109274-109283.

37. Greenland S, Follmann DA, Shih JH, Gabriel EE. Causal claims about the D-value. *Stat Probab Lett.* 2021;174:109274-109283.

38. Gail MH, Benichou J, Coyle MH, et al. Estimation of the correlation coefficient in bivariate competing risks models. *Biometrics.* 2010;66(2):334-340.

39. Gail MH, Benichou J, Coyle MH, et al. Estimation of the correlation coefficient in bivariate competing risks models. *Biometrics.* 2010;66(2):334-340.

40. Gail MH, Benichou J, Coyle MH, et al. Estimation of the correlation coefficient in bivariate competing risks models. *Biometrics.* 2010;66(2):334-340.

41. Gail MH, Benichou J, Coyle MH, et al. Estimation of the correlation coefficient in bivariate competing risks models. *Biometrics.* 2010;66(2):334-340.

42. Gail MH, Benichou J, Coyle MH, et al. Estimation of the correlation coefficient in bivariate competing risks models. *Biometrics.* 2010;66(2):334-340.
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: Dobler D, Möllenhoff K. A nonparametric relative treatment effect for direct comparisons of censored paired survival outcomes. *Statistics in Medicine*. 2024;43(11):2216-2238. doi: 10.1002/sim.10063