A consistent version of distance covariance for right-censored survival data and its application in hypothesis testing

Dominic Edelmann¹ | Thomas Welchowski² | Axel Benner¹

¹ Division of Biostatistics, German Cancer Research Center, Heidelberg, Germany
² Institute of Medical Biometry, Informatics and Epidemiology, University Hospital of Bonn, Bonn, Germany

Correspondence
Dominic Edelmann, Division of Biostatistics, German Cancer Research Center, 69120 Heidelberg, Germany.
Email: dominic.edelmann@dkfz-heidelberg.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: 417754611; German Federal Ministry of Education and Research, Grant/Award Numbers: 01ER1505a, 01ER1505b

Abstract
Distance covariance is a powerful new dependence measure that was recently introduced by Székely et al. and Székely and Rizzo. In this work, the concept of distance covariance is extended to measuring dependence between a covariate vector and a right-censored survival endpoint by establishing an estimator based on an inverse-probability-of-censoring weighted U-statistic. The consistency of the novel estimator is derived. In a large simulation study, it is shown that induced distance covariance permutation tests show a good performance in detecting various complex associations. Applying the distance covariance permutation tests on a gene expression dataset from breast cancer patients outlines its potential for biostatistical practice.

Keywords
distance correlation, distance covariance, hypothesis testing, nonlinear, survival analysis

1 | INTRODUCTION

Testing associations between continuous clinical covariates and survival are predominantly performed using some (semi-)parametric survival model, such as, for example, the Cox proportional hazards regression (Cox, 1972) or the accelerated failure time (AFT) model (Wei, 1992; Kalbfleisch and Prentice, 2011). These approaches assume that the influence of the clinical covariate on survival takes a simple parametric form. For example, a proportional hazards model with a single predictor $X$ assumes that the impact on the hazard $h(t)$ at all times $t$ is multiplicative with $h(t) = h_0(t) \exp(\beta X)$, where $h_0(\cdot)$ is the baseline hazard and $\beta \in \mathbb{R}$ is the corresponding regression coefficient.

Consequently, tests based on these models are targeted to detect parametric associations of the specified form. Notably, tests based on the Cox proportional hazards model or the AFT model have low power against alternatives that involve nonmonotone influences of covariates on survival.

Yet, the impact of many clinical parameters with survival is more complex and cannot be modeled using a simple proportional hazards model. For detecting such complex association in statistical practice, one can use visual diagnostic tools such as for example plotting the martingale residuals against the covariate values (Therneau and Grambsch, 2000). However, there is an increasing number of studies investigating the influence of high-dimensional molecular data on patient survival, see, for example Chen et al. (2007), Dvinge et al. (2013), and Ng et al. (2016). When considering such a high number of predictors—the number of covariates often exceed several thousand—visual diagnostic tools are not feasible anymore. Consequently, complex associations between high-dimensional molecular data and survival cannot be detected by the testing procedures that are predominantly used in statistical practice.
This clearly urges the need for automatic tools for finding general associations between covariates and censored data and first approaches for this purpose have been developed during the last years (Peng and Fine, 2008; Wood, 2012).

For testing dependence in uncensored data, Gábor Székely and his coauthors (Székely et al., 2007; Székely and Rizzo, 2009) recently developed the concept of distance covariance. Different from classical dependence measures such as the usual covariance, the distance covariance coefficient can detect any possible association between random variables. Since the appearance of Székely et al. (2007) and Székely and Rizzo (2009), enormous interest in the theory and applications of distance correlation has arisen; we refer to Edelmann et al. (2019) for an extensive review of distance covariance methods with a focus on time series analysis.

In some recent work, distance covariance has been applied on variable screening for high-dimensional models (Zhao and Li, 2012; Song et al., 2014; Li et al., 2016; Zhang et al., 2017; Huang et al., 2019; Pan et al., 2019). The approach in Edelmann et al. (2020) screens variables based on the distance correlation between covariates and the martingale residuals of a null Cox model, whereas Chen et al. (2018) introduce two different versions of robust distance correlation for survival data. While both approaches show good performance in simulations, neither of the two works derives an empirical version of distance correlation satisfying the desirable properties in Székely et al. (2007). Most importantly, they do not show that their estimators converge to a suitable population version of distance covariance or distance correlation and they do not derive tests, which can provably detect any dependencies between covariates and survival times.

In this work, we will fill this gap by developing an estimator for distance covariance between covariates and survival times under right-censoring. Notably, exploiting results by Datta et al. (2010) an Huo and Székely (2016), we derive the consistency and asymptotic normality of an inverse probability of censoring-weighted (IPCW) version of distance covariance. Moreover, using cumbersome combinatorial calculations, we derive computationally efficient versions of these novel estimators that can be calculated in $O(n^2)$ steps. The new estimator for distance covariance immediately induces a permutation test for detecting arbitrary associations between covariates and survival times.

The remainder of the paper is organized as follows. In Section 2, we introduce the concept of distance covariance and state the U-statistic estimate for the squared distance covariance derived by Huo and Székely (2016). Section 3 contains the definitions of the novel estimator of distance covariance for right-censored survival data and corresponding theoretical results; we also discuss important practical aspects. A simulation study in Section 4 compares a permutation test based on the estimator for distance covariance with tests based on the Cox model, thin-plate regression splines (Wood, 2003), and a test adapted from the residual distance covariance in Edelmann et al. (2020). In Section 5, the distance covariance permutation test is applied on a gene expression dataset from breast cancer patients. All proofs to theoretical results are provided in the Appendix “Proofs” in the Supporting Information.

2 | PRELIMINARIES: DISTANCE COVARIANCE

In the following, we will introduce the notion of distance covariance as defined by Székely et al. (2007) and Székely and Rizzo (2009). Moreover, we will state some important results that will be crucial for deriving the novel estimator.

Let $p$ and $q$ be positive integers. For jointly distributed random vectors $(X, Y) \in \mathbb{R}^p \times \mathbb{R}^q$ and nonrandom vectors $(s, t) \in \mathbb{R}^p \times \mathbb{R}^q$, the joint characteristic function (Van der Vaart, 2000, Section 2.3) of $(X, Y)$ is

$$\phi_{X,Y}(s, t) = \mathbb{E}\exp\left[ i(s, X) + i(t, Y) \right],$$

where $i = \sqrt{-1}$ is the imaginary unit and $\langle \cdot, \cdot \rangle$ denotes the standard scalar product. Moreover, let $\phi_X(s) = \phi_{X,Y}(s, 0)$ and $\phi_Y(t) = \phi_{X,Y}(0, t)$ be the corresponding marginal characteristic functions.

Now, denoting by $\| \cdot \|$ the standard Euclidean norm and by $| \cdot |$ the modulus in $C$, the distance covariance (Székely et al., 2007, p. 2772) between $X$ and $Y$ is the positive square root of

$$\mathcal{V}^2(X, Y) = \frac{1}{c_p c_q \pi^p \pi^q \Gamma((d+1)/2)} \int_{\mathbb{R}^{p+q}} \frac{|\phi_{X,Y}(s, t) - \phi_X(s)\phi_Y(t)|^2}{\|s\|^{p+1}\|t\|^{q+1}} ds \, dt,$$

(2.1)

where $c_d$ is the constant defined by $c_d = \pi^{(d+1)/2} / \Gamma((d+1)/2)$.

The distance covariance features several desirable properties for a dependence coefficient (Székely et al., 2007). Most importantly, $\mathcal{V}(X, Y)$ satisfies

$$\mathcal{V}(X, Y) = 0 \iff \text{if and only if } X \text{ and } Y \text{ are independent}.$$

This implies that $\mathcal{V}$ can identify any possible association between $X$ and $Y$.

Another important property of distance covariance is that it features an astonishingly simple dependence measure that only involves simple quadratic polynomials of the distances between the observations of the corresponding samples. To define this sample measure, consider that we have given samples $(X_1, Y_1), \ldots, (X_n, Y_n)$ drawn from
(X, Y) ∈ ℝ^{p+q}. Then an unbiased consistent sample version of Υ^2(X, Y) is given by (Huo and Székely, 2016)

\[ \hat{\Omega} = \frac{n}{n-3} T_1 + \frac{n^3}{(n-1)(n-2)(n-3)} T_2 - \frac{2n^2}{(n-2)(n-3)} T_3, \quad (2.2) \]

where

\[ T_1 = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \|X_i - X_j\| \|Y_i - Y_j\|, \]

\[ T_2 = \frac{1}{n^4} \left( \sum_{i=1}^{n} \sum_{j=1}^{n} \|X_i - X_j\| \right) \left( \sum_{i=1}^{n} \sum_{j=1}^{n} \|Y_i - Y_j\| \right), \]

and

\[ T_3 = \frac{1}{n^3} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \|X_i - X_j\| \|Y_i - Y_k\|. \quad (2.3) \]

Huo and Székely (2016) further show that \( \hat{\Omega} \) can be written as a U-statistic. In particular, defining the kernel function (Huo and Székely, 2016; Huang and Huo, 2017)

\[ h((X_1, Y_1), (X_2, Y_2), (X_3, Y_3), (X_4, Y_4)) = \frac{1}{4} \sum_{1 \leq i \leq 4} \|X_i - X_j\| \|Y_i - Y_j\| \]

\[ - \frac{1}{4} \sum_{1 \leq i \leq 4} \sum_{j \neq i} \|X_i - X_j\| \sum_{1 \leq j \leq 4} \|Y_i - Y_j\| \]

\[ + \frac{1}{24} \sum_{1 \leq i < j \leq 4} \sum_{1 \leq k < j \leq 4} \|X_i - X_j\| \|Y_i - Y_j\|, \quad (2.5) \]

a U-statistic representation of \( \hat{\Omega} \) is given by

\[ \hat{\Omega} = \left( \begin{array}{c} \sum_{1 \leq i < j \leq 4} \sum_{i \neq j} \sum_{s \neq i} \sum_{s < j} h((X_i, Y_i), (X_j, Y_j), (X_s, Y_s), (X_l, Y_l)) \end{array} \right)^{-1}. \quad (2.6) \]

3 METHODS

3.1 An IPCW estimator for the distance covariance under right-censoring

We consider random variables \( X \in \mathbb{R}^p, S \in [0, \infty) \) and \( C \in [0, \infty) \), where \( X \) represents a covariate vector, \( S \) represents the survival time, and \( C \) represents the censoring time. While \( X \) and \( S \) are allowed to feature any possible dependency, we assume that \( C \) is independent of \((X, S)\). Moreover, we will assume that the hazard function of the censoring distribution

\[ \lambda_C(t) = \lim_{dt \to 0} \frac{P(t \leq C < t + dt | C \geq t)}{dt} \]

is absolutely continuous. The censoring indicator will be denoted by \( \Delta = 1_{\{S \leq C\}} \) and the follow-up time by \( T = \min(S, C) \).

Given an i.i.d. sample \((X_1, T_1, \Delta_1), \ldots, (X_n, T_n, \Delta_n)\) drawn from \((X, T, \Delta)\), our goal is to detect associations between the covariate vector \( X \) and the survival times \( S \) by deriving a consistent estimator for the distance covariance \( \mathcal{V}(X, S) \). For this purpose, we will consider an IPCW versions of the U-statistic representation of distance covariance \( (2.6) \) given by

\[ \left( \begin{array}{c} \sum_{1 \leq i < j \leq 4} \sum_{i \neq j} \sum_{s \neq i} \sum_{s < j} \frac{\Delta_i}{\hat{G}_n(T_i -)} \frac{\Delta_j}{\hat{G}_n(T_j -)} \frac{\Delta_l}{\hat{G}_n(T_l -)} \end{array} \right) \times \frac{\Delta_k}{\hat{G}_n(T_k -)} h((X_i, T_i), \ldots, (X_l, T_l)). \quad (3.1) \]

Now applying (Datta et al., 2010, Theorem 1) (see Supporting Information for details) yields that (under the technical assumptions (C1)–(C3) given below), \( (3.1) \) is a consistent estimator for \( \mathcal{V}(X, S) \), where \( h(\cdot) \) denotes the kernel function defined in Equation (2.5), \( \hat{G}_n(\cdot) \) is the Kaplan–Meier estimate of \( G \), and for any function \( f \) defined on a subset of \( \mathbb{R}, f(s-) \) denote the left-sided limit of \( f(\cdot) \) in \( s \).

However, different from the empirical version of distance covariance for uncensored data defined in Equation (2.2) that can be calculated in \( O(n^2) \) steps, the representation \( (3.1) \) features a complexity of \( O(n^4) \). Consequently, this representation is computationally highly inefficient and virtually unfeasible for most practical purposes.

Using nontrivial calculations (see Appendix “Proofs” in the Supporting Information for details), we can derive an alternative representation for \( (3.1) \), that features a computational complexity of \( O(n^2) \) only. Notably, defining \( W_i = \frac{\delta_i}{\hat{G}_n(T_i -)}, a_{ij} = \|X_i - X_j\|, d_{ij} = |T_i - T_j| \), \( m := \sum_{i=1}^{n} W_i, M := \sum_{i=1}^{n} W_i^2 \), the estimator in \( (3.1) \) can be written as

\[ \tilde{\Omega} = \frac{1}{n(n-1)(n-2)(n-3)} \left( \sum_{i,j=1}^{n} W_i W_j a_{ij} d_{ij} \right) \times (m^2 - m(W_i + W_j) - M + 2W_i W_j) \]
While the representation for \( \tilde{\Omega} \) stated in (3.2) may appear rather involved on the first sight, merely elementary algebra is needed to evaluate the summands in the respective expression. Moreover, it is evident that the calculation of (3.2) can be carried out in \( O(n^2) \) steps.

For the purpose of stating the asymptotic properties of \( \tilde{\Omega} \), we introduce the notation \( Y(t) = 1_{[T \geq t]} \), \( N^C(t) = 1_{[T \leq t, \Delta = 0]} \),

\[
M^C(t) = N^C(t) - \int_0^t Y(u)\lambda_C(u)du, \\
h_1(x,t) = \mathbb{E}[h((x,t),(X_2,T_2),(X_3,T_3),(X_4,T_4))],
\]

and

\[
w(t) = \frac{1}{P(T \geq t)} \int_{\mathbb{R}^2 \times \mathbb{R}} \frac{h_1(x,u)}{G(u-)} 1_{[u>t]} dp^{(X,S,\Delta)}(x,u,1),
\]

where \( p^{(X,S,\Delta)} \) denotes the joint probability measure of the triple \((X, S, \Delta)\).

Now, assuming the conditions

(C1) \( \mathbb{E}[X]^2 < \infty, \mathbb{E}[S]^2 < \infty, \)
(C2) \( \int_{\mathbb{R}^2 \times [0,\infty)} \frac{h_1^2(x,u)}{G(u-)} p^{(X,S,\Delta)}(dx,du,1) < \infty, \)
(C3) \( \int_{[0,\infty)} w^2(t)\lambda_C(t)dt < \infty, \)

we can show (see Appendix “Proofs” in the Supporting Information) the following asymptotic results for \( \tilde{\Omega} \).

**Theorem 1.** Suppose that (C1)–(C3) hold and let

\[
\sigma^2 = 16\text{Var}\left( \frac{h_1(X,T)\Delta}{G(T-)} + \int_{[0,\infty)} w(t)dM^C(t) \right). 
\]

Then, for \( n \to \infty, \)

\[
\tilde{\Omega} \xrightarrow{P} \mathcal{N}^2(X,S).
\]

Moreover, if \( \sigma^2 > 0 \), then

\[
\sqrt{n}(\tilde{\Omega} - \mathcal{N}^2(X,S)) \xrightarrow{D} \mathcal{N}(0,\sigma^2); \\
\]

if \( \sigma^2 = 0 \), then

\[
\sqrt{n}(\tilde{\Omega} - \mathcal{N}^2(X,S)) \xrightarrow{P} 0.
\]

**Remark 1.** If \( X \) and \( Y \) are independent, it holds \( \sigma^2 = 0 \) and hence

\[
\sqrt{n}(\tilde{\Omega} \xrightarrow{P} 0).
\]

In a similar fashion, we obtain an IPCW version of distance correlation that is presented in the Supporting Information.

Before concluding this section, we briefly discuss the meaning and purpose of the conditions (C1)–(C3). The existence of the second moments of the covariates and the survival times (C1) implies the existence of the second moments of the kernel function (2.5) (see Edelmann et al. (2020)), which, in turn, guarantees the asymptotic normality of the estimator (3.2) when no censoring is present. (C2) now states that the second moments of the weighted kernel function \( \frac{h_1^2(X,T)\Delta}{G(T-)} \) exist leading to the asymptotic normality of the estimator (3.2), when the censoring distribution \( G \) is known and used for calculating the weights \( W_i \). The last condition (C3) now guarantees that this asymptotic normality is retained when \( G \) is replaced by its Kaplan–Meier estimate \( \hat{G} \); technically, it ensures that the second summand of the asymptotic variance in (3.3) is finite.

For practical purposes, it is interesting to know that, given the existence of the second moments in (C1), conditions (C2) and (C3) are satisfied if \( \lambda_C \) is bounded and the supports of \( S \) and \( C \) are of the form \([0,\tau_S]\) and \([0,\tau_C]\), respectively, with \( \tau_S < \tau_C \). This corresponds to studies, where the follow-up is longer than the maximal lifespan of an individual, but there may be censoring due to dropouts. On the other hand, (C2) is never satisfied, if the supports of \( S \) and \( C \) are of the form \([0,\tau_S]\) and \([0,\tau_C]\), respectively, with \( \tau_S > \tau_C \). This relates to the situation, where the maximal lifespan of an individual is longer than the follow-up, which is often the case in clinical studies. In these settings, variations of \( \tilde{\Omega} \) are required, which are discussed in Section 3.3.

### 3.2 Independence testing

Theorem 1 directly implies a permutation independence test between \( X \) and \( S \) given an i.i.d. sample
For each $j \in \{1, \ldots, B\}$, we then calculate the squared distance covariance based on the permuted sample $(X_{1j}, T_{1j}, \Delta_{1j}), \ldots, (X_{nj}, T_{nj}, \Delta_{nj})$, which we denote by $\tilde{\Omega}_j$. The $p$-value of the permutation test is then given by

$$\frac{\#(\{j \in \{1, \ldots, B\} \mid \tilde{\Omega}_j \geq \overline{\Omega}\}) + 1}{B + 1},$$

where for a set $P$, $\#(P)$ denotes the cardinality of $P$.

For an illustration of the computational performance, see Appendix "Computation time of the permutation test" in the Supporting Information.

### 3.3 Practical considerations

In Section 3.1, we have derived a consistent estimator for the distance covariance between covariates $X$ and the survival time $S$ under right-censoring using only mild assumptions. However, even these mild assumptions are often too restrictive in biomedical situations. Notably, it can happen that the distribution of the survival time is not completely identifiable and hence also the distance covariance between $X$ and $S$ cannot be consistently estimated. To be precise, in a clinical or observational study, the censoring time $C$ is typically upper-bounded by the total duration of the study $t_0$. On the other hand, the probability that a patient’s survival time $S$ is larger than $t_0$ is typically positive. Hence, $P(S > t_0) > 0$, which directly implies that $P(S > t \mid S > t_0)$ is not identifiable for $t > t_0$. Concerning the assumptions for Theorem 1, (C2) is obviously not satisfied in such a situation, since the corresponding integral is not finite.

As a remedy, we suggest to estimate the distance covariance of the covariate $X$ with the truncated survival times $S^\tau = \min(\tau, S)$, where $\tau \in [0, \infty)$ is a suitably chosen constant. In most medical trials, a natural choice for $\tau$ is given by the follow-up time of the trial. If this is not the case, one may consider a clinically meaningful endpoint (e.g., 5 years) or choose a point in time, where a specified number of patients is still under observation (e.g., 20%). The truncated survival time $S^\tau$ is clearly a meaningful substitute for $S$, while featuring the advantage that the distance covariance of $X$ with $S^\tau$ can be consistently estimated (as long as $EX^2 < \infty$ and $\lambda_C(t)$ is absolutely continuous). For calculating the distance covariance between $X$ and $S^\tau$, all follow-up times larger than $\tau$ are replaced by $\tau$; moreover, the corresponding censoring indicators are set to 1. We will see in Section 5 that this approach works well in practice.

Moreover, we point out that for any monotone transformation $g(\cdot)$, the distance covariance between $X$ and $g(S)$ also provides an association test that is consistent against all alternatives. In Section 5, we compare the distance covariance test after applying a log-transformation with the standard version on a real data example.

### 3.4 Computational issues in high dimensions and alternative tests

When analyzing high-dimensional covariate data, one may often perform a single association tests for each coefficient of the covariate resulting in a large number of tests. In this case, one may consider adjusting for multiple testing using, for example, a Benjamini–Hochberg or Bonferroni correction. For this purpose, it is crucial to precisely evaluate small $p$-values of the tests. When using a permutation test, this issue requires a large number of permutations and leads to an increased computation time.

For dealing with this issue, we propose two different strategies. The first strategy involves generating $B$ permuted samples and then fitting a generalized Pareto distribution to the tail of the empirical distribution of the corresponding statistics $\tilde{\Omega}_1, \ldots, \tilde{\Omega}_B$; subsequently, the $p$-values are obtained by the fitted distribution.

The second strategy is a stepwise procedure. In the first step, the $p$-value for each test is calculated using a small number of permutations. In the second step, only tests for which $p$-values fall below a certain threshold are reevaluated using a larger number of permutations. In subsequent steps, very small $p$-values can then be more precisely evaluated using even more permutations.

More explanation and a comparison of the two different strategies are provided in the Supporting Information.

### 4 SIMULATION STUDY

To evaluate the performance of the proposed distance-likelihood ratio covariance permutation test, a simulation study was conducted. The novel method was compared to likelihood ratio (LR) tests based on the AFT and the Cox proportional hazards models with a linear predictor. As competitors that take into account nonlinear effects, we considered a Cox proportional hazards model with a thin-plate regression spline (Wood, 2003), a marginal test between covariates and martingale residuals of an empty Cox model based on sliced average variance estimation (Shao et al., 2007) and an ad-hoc permutation test based on the distance covariance between covariates and martingale residuals, cf. Edelmann et al. (2020). For the thin-plate regression spline, a Wald-type test was used to test the null
hypothesis that the smooth term equals zero (Wood, 2012). The regression spline was set to five degrees of freedom with one covariate and 25 degrees of freedom in the case of two covariates. The number of permutations in the two permutation tests was set to $B = 1000$.

The simulation study consists of two phases. The first phase uses prior simulations to specify the simulation parameters for the second phase. In the second phase, the different methods are compared with respect to power and type I error rates. For the first phase, a large dataset $D_1$ of $N = 10^6$ observations was randomly drawn once. For each observation, three covariates were generated independently with uniform distribution $\forall j = 1, \ldots, 3 : X_j \sim U(-3, 3)$. Based on the dataset $D_1$, nine different linear predictor $\eta$ structures with one covariate and five two-dimensional $\eta$ structures were calculated. Explicit formulas used in the construction of $\eta$ are given in Appendix “Simulation Study” in the Supporting Information; a graphical illustration of the different structures is given in Figure 1. This figure appears in color in the electronic version of this article, and any mention of color refers to that version. We investigated, for example, polynomials (linear, quadratic, cubic), cyclical patterns (cosinus, modulus, Rastrigin), and functions with random components (twoLines, twoParabolas, condVar). Survival and censoring times were log-normally distributed with $T \sim \log\text{-Normal}(\mu_{T, \log} = \eta, \sigma_{T, \log})$ and $C \sim \log\text{-Normal}(\mu_{C, \log}, \sigma_{C, \log})$. Using the generated dataset $D_1$, $\eta$ was standardized. For each $\eta$, the standard deviation $\sigma_{T, \log}$ was set to achieve prespecified signal to noise ratios. Linear, quadratic, cubic, and piecewise constant scenarios signal-to-noise ratios were set to 0.125, cosinus and circle to 0.25, two lines and parabolas as well as two-way interactions to 0.5, and all remaining $\eta$ structures to 0.75. For each $\eta$, the parameters $\mu_{C, \log}$ and $\sigma_{C, \log}$ were then set to three different pairs of values corresponding to censoring rates of 0.25, 0.5, and 0.75, respectively. Finally, for each $\eta$ structure, $\tau$ was set to the 90% quantile of the survival times. All prior parameters (standardization of $\eta$, $\sigma_{T, \log}$, $\sigma_{C, \log}$, $\tau$) were computed once for each $\eta$ structure and then remain fixed for the second phase of the simulation procedure.

In the second phase, the different tests were compared in terms of type I error and power. For each sample size $n \in \{100, 200, \ldots, 1000\}$ and each of the 14 $\eta$ structures, we simulated $k = 1000$ datasets, corresponding to $n$ observations of the vector $(X_1, X_2, X_3, T, C) \in \mathbb{R}^5$. All survival times were cut by the parameter $\tau$ (corresponding to the 90% quantile of the survival times in the dataset $D_1$). We evaluated six different tests: AFT LR test, Cox proportional hazards with linear effects (Cox) LR test, Cox proportional hazards with thin-plate regression spline (CoxSpline) Wald test, permutation test based on the IPCW distance covariance ($DCov$), test based on sliced average variance estimation ($SAVE_{\text{martingale}}$), and permutation test based on the distance covariance between covariates and martingale residuals ($Dcov_{\text{martingale}}$). Power was evaluated by testing the influential covariates with survival. Type I error rates were estimated by evaluation of the test statistics based on the third noise covariate, which had no relation to the survival time.

The results of the type I error rates are provided in Figures 3–5 in Appendix “Simulation Study” of the Supporting Information. In all scenarios, the type I errors of all tests are close to the nominal threshold $\alpha = .05$. In most cases, the AFT model test has slightly higher type I error than the Cox model based test (Cox). The type I error rates of the IPCW distance covariance test ($DCov$) were quite comparable to those of the martingale-based distance covariance test ($Dcov_{\text{martingale}}$) and the Wald test based on thin-plate regression spline (CoxSpline). The $SAVE_{\text{martingale}}$ methodology showed slightly higher type I errors.

The results of the power analyses are given in Figures 2–4. Most considered tests showed a high power in discovering linear, quadratic, and cubic effects. Even AFT and Cox could detect quadratic and cubic effects, which is likely due to the positive trend in the considered $\eta$ structure. Only $SAVE_{\text{martingale}}$ showed a substantially worse performance than all other tests in detecting linear and cubic relationships. The two distance covariance based tests give very similar results and show a very good overall performance. In the low censoring scenario, $DCov$ shows the highest power in four scenarios (cosine, condVar, modulus, shubert). $CoxSpline$ has the highest power in three scenarios (piecewise constant, circle, rastrigin), and $SAVE_{\text{martingale}}$ has the highest power in two scenarios (twoLines, twoParabolas). As expected, tests based on linear models (AFT, Cox) could not detect the two-way interaction and continuous XOR example. $DCov$, $Dcov_{\text{martingale}}$, $SAVE_{\text{martingale}}$, and $CoxSpline$ performed equally well in those cases. The results for moderate censoring are similar; however, $DCov$ seems to show some slight problems with heavy censoring.

5 | REAL DATA EXAMPLE

For demonstrating the performance of the proposed distance covariance test on real data, we applied this method on the breast cancer microarray dataset investigated by Van De Vijver et al. (2002). The dataset consists of expression values of 24,481 genes in 295 tumor samples from different patients. Moreover, it also contains data on overall survival for each individual and additional patient data such as estrogen receptor status and age. The median
FIGURE 1  Visualization of all relationships of $\eta$ used in the simulation study. The upper part displays univariate and the lower part bivariate relationships. The magnitude of $\eta$ for bivariate relationships is displayed in the bottom right corner. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.
follow-up time of the study was 8.85 years, the event was observed for 79 patients.

To search for genes that are associated with overall survival, we applied both an LR test based on a standard Cox model and the proposed distance covariance permutation test using \( B = 10,000 \) permutations. For the distance covariance test, we restricted our dataset to a 10-year follow-up. The censoring indicator of all individuals surviving 10 years was set to 1 and its survival time to 10 years; this corresponds to considering the truncated endpoint \( \min(S, 10\text{ years}) \), see Section 3.3 for details.

For comparison of the two tests, we first evaluated the number of rejections for association tests between single genes and overall survival using a gene-wise significance level of \( \alpha = 0.05 \). The number of rejections for the distance covariance permutation test and the LR test based on the Cox model are illustrated on the left-hand side of Figure 5. Additionally to the test controlling the gene-wise significance level of .05, we additionally applied a Benjamini–Hochberg (BH) correction to control for a false-discovery rate (FDR) of 0.1. The number of rejections for an FDR of 0.1 is illustrated on the right-hand side of Figure 5.

When controlling for a gene-wise significance level for .05, the distance covariance test leads to 6649 rejections, whereas the tests based on the Cox model reject the null hypothesis of no association in 5362 cases. When adjusting the \( p \)-values using the Benjamini–Hochberg procedure and controlling for an FDR of 0.1, we obtain 4417 rejections for the distance covariance test and 3107 rejections for the Cox model. In total, these results suggest that, for this example, the distance covariance test is superior in detecting associations between gene expression and survival.

Additionally, we also performed the distance covariance permutation test on log-transformed follow-up times to investigate possible benefits arising from the use of such a transformation. When controlling for a gene-wise significance level for .05, the test on log-transformed survival times leads to 6759 rejections. When adjusting the \( p \)-values using the Benjamini–Hochberg procedure and controlling for an FDR of 0.1, the null hypothesis of no association is rejected in 4633 cases. A corresponding Venn diagram is provided in the Appendix “Real Data Example” in the Supporting Information. In total, we observe a slightly higher number of rejections when applying a log-transformation, suggesting that it may be beneficial to apply such a transformation in practice.

To better understand the nature of the dependencies, which are detected by a distance covariance test, but not by a Cox model, we performed an exploratory analysis investigating the associations that turned out to be highly
significant ($p$-value < $10^{-3}$) for the distance covariance test, but nonsignificant ($p$-value > .05) for the Cox model-based test. In total, there were 101 of such associations. First, we fitted Cox models using penalized splines via the `pspline` function from the R package `survival`, which revealed nonlinear functional dependencies (i.e., relationships, where the hazard can be expressed by a function of the gene) between numerous genes and the corresponding log-hazard ratio (vs. a reference level of gene expression 0). Four of these dependencies are displayed in Figure 6. The upper two plots illustrate genes with a U-shaped and reverse U-shaped relationship with the log-hazard ratio, respectively. In the lower left plot, the log-hazard ratio seems to be approximately constant for expression values smaller than 0.2, followed by a sharp decrease. Finally the hazard ratio in the lower right plot is decreasing for small expression values, approximately constant for intermediate values and increasing for large expression values.

However, as we have seen in Section 4, distance covariance cannot only detect functional but any kind of relationships between a predictor and its log-hazard ratio, see the examples of two lines or two parabolas in Section 4. To investigate nonfunctional relationships, we fitted separate standard Cox models on different subgroups of the data. As an example, we considered the subgroups of patients with a small tumor (diameter $\leq$ 20 mm, $n = 155$ patients) and patients with a large tumor (diameter $> 20$ mm, $n = 140$ patients). We also explored the subgroups of younger (age $\leq$ 40 years) and older (age $> 40$ years) women. In 46 of the 101 genes under consideration, a standard Cox model in at least one of the four subgroups gave a significant result. While this could possible indicate that there is a subgroup effect for some of these genes, we also emphasize that we encounter multiple testing issues at this point. In total, we found the evidence of subgroup effects in these four subgroups rather unconvincing.

Finally, since the distance standard deviation is more robust against outliers and heavy-tailed distributions than the classical standard deviation (Edelmann et al., 2020), one may conjecture that the distance covariance independence test for survival is more robust against outliers than the Cox model. Indeed, when removing the four most extreme observations (in terms of distance to the median) from each gene, Cox model based tests gives significant results in 47 of the 101 genes under consideration; this observation suggests that the distance covariance test may indeed be more robust than the LR test based on a standard Cox model.
FIGURE 4  Estimated power of six different tests over all η structures and varying sample sizes. Heavy censoring (censoring rate 0.75). This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

| Test Type       | Sample Size |
|-----------------|-------------|
| linear          | 250         |
| quadratic       | 500         |
| cubic           | 750         |
| cosine          | 1000        |
| twoLines        | 250         |
| twoParabolas    | 500         |
| pieceConst      | 750         |
| modulus         | 1000        |
| condVar         |             |
| circle          |             |
| twoWayInter     |             |
| XOR             |             |

FIGURE 5  Number of rejections for association tests between single genes and overall survival for the distance covariance permutation test and the likelihood ratio test based on the Cox model. The left-hand side illustrates the number of rejections for a gene-wise significance level of 0.05. The right-hand side gives the number of rejections controlling for a false discovery rate (FDR) of 0.1 using Benjamini–Hochberg (BH) adjustment of the p-values.

| Test Type       | Number of Rejections |
|-----------------|-----------------------|
| Distance Covariance | 2634                  |
| Cox model       | 2438                  |
| SAVE_martingale | 669                   |
| CoxSpline       | 1347                  |
| Dcov            | 4015                  |
| Dcov_martingale |                      |

| Test Type       | Number of Rejections |
|-----------------|-----------------------|
| Distance Covariance | 1979                  |
| Cox model       | 1347                  |
| SAVE_martingale | 669                   |
| CoxSpline       | 4015                  |
| Dcov            | 2438                  |
| Dcov_martingale |                      |
FIGURE 6 Association between the expression values of four different genes (Contig7558_RC, NM_006624, NM_002359, and AF248651) and the corresponding log-hazard ratios for overall survival

6 | DISCUSSION

The number of studies investigating associations between survival and high-dimensional molecular data has been rapidly increasing throughout the last decade. Concurrent with this development, a large number of different methods for testing and modeling in high-dimensional survival data have been developed (Tibshirani, 1997; Goeman et al., 2005; Gorst-Rasmussen and Scheike, 2013). An aspect, which is often neglected in hypothesis testing, is the possible presence of nonmonotone effects. Consequently, the overwhelming majority of approaches have little power against these alternatives.

In this work, we have extended the well-studied distance correlation approach to right-censored survival data. Notably, we have established consistent estimators for the distance covariance and distance correlation between a covariate vector and the survival time. Using permutation tests based on the distance covariance estimator allows for testing against general alternatives in this setting. Both simulation results and a real data example demonstrate that this test appears to be a very promising approach, outperforming existing methods in numerous settings. Especially the real data example shows that this approach can detect a considerable number of associations that a standard Cox model would miss.

This article may be seen as a first step into distance covariance testing for survival data. Numerous extensions of the distance correlation and distance covariance methodology have been developed and it is obvious that many of these extensions are also sensible in the survival setting. An important topic is, for example, the consideration of confounders using partial (Székely and Rizzo, 2014) or conditional (Wang et al., 2015) distance covariance. Moreover, while permutation tests are still the standard option for distance covariance in applications, parametric tests have recently been developed, which may further speed up the necessary computations (Berschneider and Böttcher, 2018). Finally, distance covariance methods are not restricted to covariates in \( \mathbb{R}^p \); the concept of generalized distance covariance (Lyons, 2013; Sejdinovic et al., 2013) actually allows for testing independence between random variables on general metric spaces. It is our goal to address extensions of these methods on survival data in our future work.

ACKNOWLEDGMENTS

We thank Thomas Alexander Gerds and Paul Blanche for helpful discussions. Dominic Edelmann is funded by the Deutsche Forschungsgemeinschaft (DFG), Project No. 417754611. This work was supported by grants from
the German Federal Ministry of Education and Research (BMBF) (grant numbers: 01ER1505a and 01ER1505b).

**DATA AVAILABILITY STATEMENT**

The data that support the findings of this paper are openly available in the R Package cancerdata at https://bioconductor.org/packages/release/data/experiment/html/cancerdata.html [10.18129/B9.bioc.cancerdata] by Budczies and Kosztyla (2020).

**ORCID**

Dominic Edelmann https://orcid.org/0000-0001-7467-6343

**REFERENCES**

Berschneider, G. and Böttcher, B. (2018) On complex Gaussian random fields, Gaussian quadratic forms and sample distance multivariate. arXiv preprint arXiv:1808.07280.

Budczies, J. and Kosztyla, D. (2020) cancerdata: Development and validation of diagnostic tests from high-dimensional molecular data: Datasets. R package version 1.28.0, https://doi.org/10.18129/B9.bioc.cancerdata. Accessed December 23, 2020.

Chen, H.-Y., Yu, S.-L., Chen, C.-H., Chang, G.-C., Chen, C.-Y., Yuan, A et al. (2007) A five-gene signature and clinical outcome in non-small-cell lung cancer. *New England Journal of Medicine*, 356, 11–20.

Chen, X., Chen, X. and Wang, H. (2018) Robust feature screening for ultra-high dimensional right censored data via distance correlation. *Computational Statistics & Data Analysis*, 119, 118–138.

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

Datta, S., Bandyopadhyay, D. and Satten, G.A. (2010) Inverse probability of censoring weighted U-statistics for right-censored data with an application to testing hypotheses. *Scandinavian Journal of Statistics*, 37, 680–700.

Dvinge, H., Git, A., Grif, S., Salmon-Divon, M., Curtis, C., Sottoriva, A et al. (2013) The shaping and functional consequences of the microrna landscape in breast cancer. *Nature*, 497, 378–382.

Edelmann, D., Fokianos, K. and Pittsillou, M. (2019) An updated literature review of distance correlation and its applications to time series. *International Statistical Review*, 87, 237–262.

Edelmann, D., Hummel, M., Hielscher, T., Saadati, M. and Benner, A. (2020) Marginal variable screening for survival endpoints. *Biometrical Journal*, 62, 610–626.

Edelmann, D., Richards, D. and Vogel, D. (2020) The distance standard deviation. *The Annals of Statistics*, 48, 3395–3416.

Goeman, J.J., Oosting, J., Clelon-Jansen, A.-M., Anninga, J.K. and Van Houwelingen, H.C. (2005) Testing association of a pathway with survival using gene expression data. *Bioinformatics*, 21, 1950–1957.

Gorst-Rasmussen, A. and Scheike, T. (2013) Independent screening for single-index hazard rate models with ultrahigh dimensional features. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75, 217–245.

Huang, C. and Huo, X. (2017) A statistically and numerically efficient independence test based on random projections and distance covariance. arXiv preprint arXiv:1701.06054.

Huang, T.-J., McKeague, I.W. and Qian, M. (2019) Marginal screening for high-dimensional predictors of survival outcomes. *Statistica Sinica*, 29, 2105–2139.

Huo, X. and Székely, G.J. (2016) Fast computing for distance covariance. *Technometrics*, 58, 435–447.

Kalbfleisch, J.D. and Prentice, R.L. (2011) *The Statistical Analysis of Failure Time Data*. Hoboken, NJ: John Wiley & Sons.

Li, J., Zheng, Q., Peng, L. and Huang, Z. (2016) Survival impact index and ultrahigh-dimensional model-free screening with survival outcomes. *Biometrics*, 72, 1145–1154.

Lyons, R. (2013) Distance covariance in metric spaces. *The Annals of Probability*, 41, 3284–3305.

Ng, S.W., Mitchell, A., Kennedy, J.A., Chen, W.C., McLeod, J., Ibrahimova, N et al. (2016) A 17-gene stemness score for rapid determination of risk in acute leukaemia. *Nature*, 540, 433–437.

Pan, W., Wang, X., Xiao, W. and Zhu, H. (2019) A generic sure independence screening procedure. *Journal of the American Statistical Association*, 114, 928–937.

Peng, L. and Fine, J. (2008) Nonparametric tests for continuous covariate effects with multistate survival data. *Biometrics*, 64, 1080–1089.

Sejdinovic, D., Sriperumbudur, B., Gretton, A. and Fukumizu, K. (2013) Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41, 2263–2291.

Shao, Y., Cook, R.D. and Weihsberg, S. (2007) Marginal tests with sliced average variance estimation. *Biometrika*, 94, 285–296.

Song, R., Lu, W., Ma, S. and Jessie Jeng, X. (2014) Censored rank independence screening for high-dimensional survival data. *Biometrika*, 101, 799–814.

Székely, G.J. and Rizzo, M.L. (2009) Brownian distance covariance. *The Annals of Applied Statistics*, 3, 1236–1265.

Székely, G.J. and Rizzo, M.L. (2014) Partial distance correlation with methods for dissimilarities. *The Annals of Statistics*, 42, 2382–2412.

Székely, G.J., Rizzo, M.L. and Bakirov, N.K. (2007) Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35, 2769–2794.

Therneau, T.M. and Grambsch, P.M. (2000) *Modeling survival data: Extending the Cox model*. New York; London: Springer.

Tibshirani, R. (1997) The lasso method for variable selection in the Cox model. *Statistics in Medicine*, 16, 385–395.

Van De Vijver, M.J., He, Y.D., Van’t Veer, L.J., Dai, H., Hart, A.A., Voskuil, D.W et al. (2002) A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, 347, 1999–2009.

Van der Vaart, A.W. (2000) *Asymptotic Statistics*. Cambridge, MA: Cambridge University Press.

Wang, X., Pan, W., Hu, W., Tian, Y. and Zhang, H. (2015) Conditional distance correlation. *Journal of the American Statistical Association*, 110, 1726–1734.

Wei, L.-J. (1992) The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. *Statistics in Medicine*, 11, 1871–1879.

Wood, S.N. (2003) Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65, 95–114.

Wood, S.N. (2012) On p-values for smooth components of an extended generalized additive model. *Biometrika*, 100, 221–228.

Zhang, J., Liu, Y. and Wu, Y. (2017) Correlation rank screening for ultrahigh-dimensional survival data. *Computational Statistics & Data Analysis*, 108, 121–132.
Zhao, S.D. and Li, Y. (2012) Principled sure independence screening for cox models with ultra-high-dimensional covariates. Journal of Multivariate Analysis, 105, 397–411.

SUPPORTING INFORMATION
Web Appendices, Tables, and Figures referenced in Sections 3.4 and 5 are available with this paper at the Biometrics website on Wiley Online Library. R implementations of the IPCW distance covariance estimator, distance correlation estimator and the corresponding permutation test are provided in the R package dcor tools available from https://github.com/edelmand21/dcor tools.

How to cite this article: Edelmann, D., Welchowski, T., Benner, A. (2022) A consistent version of distance covariance for right-censored survival data and its application in hypothesis testing. Biometrics, 78, 867–879. https://doi.org/10.1111/biom.13470