Statistica Sinica **29** (2019), 2105-2139
doi:https://doi.org/10.5705/ss.202017.0298

MARGINAL SCREENING FOR HIGH-DIMENSIONAL PREDICTORS OF SURVIVAL OUTCOMES

Tzu-Jung Huang, Ian W. McKeague and Min Qian

Columbia University

Abstract: This study develops a marginal screening test to detect the presence of significant predictors for a right-censored time-to-event outcome under a high-dimensional accelerated failure time (AFT) model. Establishing a rigorous screening test in this setting is challenging, because of the right censoring and the post-selection inference. In the latter case, an implicit variable selection step needs to be included to avoid inflating the Type-I error. A prior study solved this problem by constructing an adaptive resampling test under an ordinary linear regression. To accommodate right censoring, we develop a new approach based on a maximally selected Koul–Susarla–Van Ryzin estimator from a marginal AFT working model. A regularized bootstrap method is used to calibrate the test. Our test is more powerful and less conservative than both a Bonferroni correction of the marginal tests and other competing methods. The proposed method is evaluated in simulation studies and applied to two real data sets.

Key words and phrases: Accelerated failure time model, bootstrap, family-wise error rate, inverse probability weighting, multiple testing, post-selection inference.

1. Introduction

The problem of detecting informative predictors of a survival outcome has received much attention over the past decade, especially since the advent of high-throughput genomic data. For example, a specific gene expression may influence a patient's survival time from diffuse large B-cell lymphoma (DLBCL). Identifying such associations from massive collections of gene-expression data remains a challenging issue. Motivated by a DLBCL study (Rosenwald et al. (2002)), we consider the fundamental detection problem of whether there exists at least one predictor (or genetic feature) that is associated with the survival outcome in the presence of right censoring.

To address this problem, we develop an adaptive resampling test for survival data (ARTS), related to the approach developed by McKeague and Qian (2015) (henceforth, MQ) for uncensored outcomes. This test provides marginal screening of the predictors, along with rigorous control of the family-wise error rate

(FWER) resulting from the implicit multiple testing. Furthermore, our testing procedure adjusts for low-dimensional baseline clinical covariates that are not included in the systematic screening of the gene-expression measurements. To identify the full set of active predictors, we propose a forward-stepwise version of the ARTS procedure that adjusts for previously included predictors at each step, and continues until no further significant predictors are found.

We specify the link between the survival outcome and the predictors in terms of a general semiparametric accelerated failure time (AFT) model that does not make any distributional assumption on the error term. Our approach also applies when the error distribution is modeled parametrically (as in Kalbfleisch and Prentice (2002), Medeiros et al. (2014)), although we focus on the semiparametric case. Let T be the (log-transformed) time-to-event outcome, and $\boldsymbol{U} = (U_1, \ldots, U_p)^T$ denote a p-dimensional vector of predictors. Here, p can be large, although it is taken to be fixed for the purpose of developing the asymptotic theory. The AFT model is given by

$$T = \alpha_0 + \boldsymbol{U}^T \boldsymbol{\beta}_0 + \varepsilon, \tag{1.1}$$

where $\alpha_0 \in \mathbb{R}$ is an intercept, and $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is a vector of regression coefficients. We assume that the error term ε has a zero mean and finite variance, and is uncorrelated with $\boldsymbol{U}$. The transformed survival outcome T is possibly right-censored by C, which is assumed to be independent of $(T, \boldsymbol{U})$ and bounded above by τ, the time to the end of the follow-up. We also make the standard assumption that $P(T \leq C) > 0$ to ensure that sufficient failure times are observed over the follow-up period (asymptotically).

In the framework of semiparametric AFT models, Koul, Susarla and Van Ryzin (1981) (henceforth, KSV) introduced the technique of inversely weighting the observed outcomes by the Kaplan–Meier estimate for the censoring, enabling them to apply standard least squares estimators from the uncensored linear model. Subsequently, two additional sophisticated methods were proposed to fit the semiparametric AFT model. The Buckley–James estimator replaces the censored survival outcome by the conditional expectation of T, given the data (Buckley and James (1979), Ritov (1990)). The rank-based method is an estimating equation approach formulated in terms of the partial likelihood score function (Tsiatis (1990), Lai and Ying (1991a), Lai and Ying (1991b), Ying (1993), Jin et al. (2003)). Our proposed marginal screening test is based on the KSV estimator, which has an advantage over the Buckley–James and rank-based methods in that it preserves a direct link with the linear model. In particular,

it maintains the marginal correlations between the inversely weighted response and the predictors.

An especially attractive feature of the AFT model is that the marginal association between T and each predictor can be represented directly in terms of a correlation. As discussed below, this allows us to reduce the high-dimensional screening problem to a single test of whether the most correlated predictor with T is significant. The most popular approach to the screening of predictors in survival analyses is to use relative or excess conditional hazard function representations of associations. However, the AFT approach has the advantage that a lack of any marginal correlation implies the absence of all correlation between T and $\boldsymbol{U}$; in the hazard-rate setting, there is no such connection.

Another attractive feature of the AFT model is that it is relatively insensitive to unmeasured heterogeneity, because the error term can act as a latent variable representing omitted confounders (Keiding, Andersen and Klein (1997)). In hazard-rate approaches, latent variables are typically included using inflexible parametric frailty models that are not easily applied in practice. In general, the presence of unmeasured heterogeneity causes the attenuation of parameter estimates. This is especially pronounced in hazard-rate approaches, such as the Cox model or additive risk models (Lin and Ying (1994), McKeague and Sasieni (1994)). On the other hand, such attenuation is much less problematic for the AFT model because the error term is only assumed to be uncorrelated with the predictors, and requires no special distributional assumption.

Under the AFT model (1.1), we test the null hypothesis $\boldsymbol{\beta}_0 = 0$, that is, that no predictor is linearly associated with T, against the omnibus alternative. The data consist of independent and identically distributed (i.i.d.) copies $(X_i, \delta_i, \boldsymbol{U}_i)$, for $i = 1, \ldots, n$, of $(X, \delta, \boldsymbol{U})$, where $X = \min(T, C)$ and $\delta = 1(T \leq C)$. The ARTS marginal screening procedure fits a series of working AFT models using one component of $\boldsymbol{U}$ at a time, and then selects the marginal KSV regression parameter estimate $\hat{\theta}_n$ that has the maximal absolute value. When the predictors are pre-standardized, the maximal regression parameter corresponds to the maximal correlation between T and any component of $\boldsymbol{U}$, motivating $\sqrt{n}\hat{\theta}_n$ as a suitable test statistic. The limiting distribution of this test statistic is nonregular (discontinuous at zero as a function of $\boldsymbol{\beta}_0$), making it difficult to calibrate the test, as explained in the standard linear regression setting by MQ. Furthermore, the presence of censoring introduces additional (discontinuous) dispersion in the limiting distribution of $\sqrt{n}\hat{\theta}_n$, which needs to be addressed.

The marginal KSV estimates stem from regressing the estimated synthetic

response $Y = \delta X/\hat{G}_n(X)$ on successive components of $\boldsymbol{U}$, where Y is regarded as an inverse probability weighted estimate, and $\hat{G}_n$ is the standard Kaplan–Meier estimator of the survival function of C (denoted by G_0). Under independent censoring (as stated earlier), the use of least squares estimators, treating Y as a response variable, is justified in view of the uniform consistency of $\hat{G}_n$ under mild conditions (e.g., when the distribution functions of T and C have no common jumps; see Stute and Wang (1993)). Independent censoring is a common assumption in the high-dimensional screening of predictors for survival outcomes (He, Wang and Hong (2013), Song et al. (2014), Li et al. (2016)). However, it is much less restrictive to assume that T and C are conditionally independent, given $\boldsymbol{U}$, in which case the conditional survival function $G_0(\cdot|\boldsymbol{U})$ of C given $\boldsymbol{U}$ can depend on the predictors. Estimating $G_0(\cdot|\boldsymbol{U})$ is challenging unless there is prior knowledge that only a single predictor is involved, using a local Kaplan–Meier estimator (Dabrowska (1989)). For simplicity, however, we assume independent censoring throughout.

Variable selection methods for right-censored survival data are widely available, although formal testing procedures are far less prevalent. For example, variants of the regularized Cox regression have been studied by Tibshirani (1997), Fan and Li (2002), Bunea and McKeague (2005), Zhang and Lu (2007), Bøvelstad, Nygård and Borgan (2009), Engler and Li (2009), Antoniadis, Fryzlewicz and Letué (2010), Binder, Porzelius and Schumacher (2011), Wu (2012), and Sinnott and Cai (2016). Penalized AFT models have been considered by Huang, Ma and Xie (2006), Datta, Le-Rademacher and Datta (2007), Johnson (2008), Johnson, Lin and Zeng (2008), Cai, Huang and Tian (2009), Huang and Ma (2010), Bradic, Fan and Jiang (2011), Ma and Du (2012), and Li, Dicker and Zhao (2014). These methods ensure the consistency of variable selection only (i.e., the oracle property), and do not address the issue of post-selection inference. Fang, Ning and Liu (2017) have established asymptotically valid confidence intervals for a preconceived regression parameter in a high-dimensional Cox model after variable selection on the remaining predictors, but this does not apply to marginal screening (where no regression parameter is singled out, a priori). Zhong, Hu and Li (2015) have considered the same problem for preconceived regression parameters within a high-dimensional additive risk model. Taylor and Tibshirani (2018) recently proposed a method of finding post-selection corrected p-values and confidence intervals for the Cox model based on conditional testing. However, to the best of our knowledge, their method has not been explored theoretically (except in a linear regression setting with independent normal errors; see Lockhart et al.

(2014)).

Statistical methods for variable selection based on marginal screening on survival data have been studied by Fan, Feng and Wu (2010), who extended sure independence screening to survival outcomes based on the Cox model. Their method applies to the selection of components of ultra-high-dimensional predictors, although no formal testing is available. Other relevant references include Zhao and Li (2012), Gorst-Rasmussen and Scheike (2013), He, Wang and Hong (2013), Song et al. (2014), Zhao and Li (2014), Hong, Kang and Li (2018), Li et al. (2016), and Hong et al. (2018).

The remainder of the paper is organized as follows. In Section 2, we formulate the testing problem and introduce the proposed test statistic based on marginal KSV estimators. The adaptive bootstrap procedure used to calibrate the test is provided at the end of Section 2. In Section 3, we propose a variant of ARTS that adjusts for the effect of baseline clinical covariates. A forward-stepwise ARTS procedure is developed in Section 4. Various competing methods are discussed in Section 5. The numerical results reported in Section 6 show that ARTS performs favorably compared with these competing methods. In Section 7, we present applications to gene-expression data and primary biliary cirrhosis data. Concluding remarks are given in Section 8. The proofs of all the results are provided in the online Supplementary Material.

2. ARTS Procedure

2.1. Preliminaries

The method proposed by Koul, Susarla and Van Ryzin (1981) for fitting the AFT model (1.1) replaces T by the synthetic response $\tilde{Y} = \delta X/G_0(X)$, which is justified by the property

$$E[\tilde{Y}|\boldsymbol{U}] = E\left[\frac{\delta X}{G_0(X)} \mid \boldsymbol{U}\right] = E\left[\frac{T}{G_0(T)} E\left[\delta|T\right] \mid \boldsymbol{U}\right] = E[T|\boldsymbol{U}], \tag{2.1}$$

where G_0 is unknown, but can be estimated by its Kaplan–Meier estimator. In other words, T and $\tilde{Y}$ have identical conditional means, given $\boldsymbol{U}$, assuming independent censoring. Therefore, we can recast the AFT model as $\tilde{Y} = \alpha_0 + \boldsymbol{U}^T\boldsymbol{\beta}_0 + \tilde{\varepsilon}$, using a new error term $\tilde{\varepsilon}$ that still has a zero mean and finite variance, and is uncorrelated with $\boldsymbol{U}$ (see the Supplementary Material for a detailed proof). Using similar arguments, we can show that $E[\tilde{Y}^2] = E[T^2/G_0(T)] \geq E[T^2]$ and $E[U_j\tilde{Y}] = E[U_jT]$, for $j = 1, \ldots, p$. Hence, this property implies that the correlation between T and U_j is uniformly proportional to the correlation between

$\tilde{Y}$ and U_j over j, leading to the equality

$$\arg \max_{j=1,\ldots,p} |\mathrm{Corr}(U_j, T)| = \arg \max_{j=1,\ldots,p} |\mathrm{Corr}(U_j, \tilde{Y})|. \tag{2.2}$$

In the next section, we use (2.2) to reduce the screening problem to a test of whether the most correlated predictor with T (or, equivalently, with $\tilde{Y}$) is significant. In practice, we recommend the pre-standardization of the predictors (as is common in variable selection) to provide scale-invariance. However, we develop the ARTS procedure in terms of the unstandardized predictors for simplicity of notation.

2.2. Maximally selected KSV estimator

To specify the predictor that is the most correlated with T, we introduce the notation

$$j(\boldsymbol{b}) = \arg \max_{j=1,\ldots,p} |\mathrm{Corr}(U_j, \boldsymbol{U}^T\boldsymbol{b})| \text{ for any } \boldsymbol{b} \in \mathbb{R}^p. \tag{2.3}$$

Under model (1.1), it is natural to have $\mathrm{Corr}(U_j, T) = \mathrm{Corr}(U_j, \boldsymbol{U}^T\boldsymbol{\beta}_0)$, which indicates that $j(\boldsymbol{\beta}_0) = \arg \max_{j=1,\ldots,p} |\mathrm{Corr}(U_j, T)|$. We assume $j(\boldsymbol{\beta}_0)$ is unique when $\boldsymbol{\beta}_0 \neq \mathbf{0}$. Thus, testing whether $\boldsymbol{\beta}_0 = \mathbf{0}$ is equivalent to a test of

$$H_0 : \theta_0 = 0 \quad \text{versus} \quad H_A : \theta_0 \neq 0,$$

where θ_0 denotes the marginal regression coefficient of $U_{j(\boldsymbol{\beta}_0)}$, the most correlated predictor with T (or, equivalently, with $\tilde{Y}$ by (2.2)). Henceforth, for notational simplicity, we denote the label $j(\boldsymbol{\beta}_0)$ by j_0.

The synthetic response $\tilde{Y}$ is not observed, but it can be estimated by $Y = \delta X/\hat{G}_n(X)$, which leads to the sample version of j_0 given by

$$\hat{j}_n = \arg \max_{j=1,\ldots,p} \left| \frac{\mathbb{P}_n(U_j - \mathbb{P}_n U_j)Y}{S_j S_Y} \right|, \tag{2.4}$$

where $\mathbb{P}_n$ is the empirical distribution, and S_j and S_Y are the sample standard deviations of U_j and Y, respectively. The best fitting marginal linear model for T with predictor U_{j_0} has the intercept and slope

$$(a_0, \theta_0) = \left(ET - \theta_0 E U_{j_0}, \frac{\mathrm{Cov}(U_{j_0}, T)}{\mathrm{Var}(U_{j_0})} \right).$$

The maximally selected KSV estimator of (a_0, θ_0) is

$$(\hat{\alpha}_n, \hat{\theta}_n) = \left(\mathbb{P}_n Y - \hat{\theta}_n \mathbb{P}_n U_{\hat{j}_n}, \frac{1}{S^2_{\hat{j}_n}} \mathbb{P}_n (U_{\hat{j}_n} - \mathbb{P}_n U_{\hat{j}_n}) Y \right), \tag{2.5}$$

where $S^2_{\hat{j}_n}$ denotes the sample variance of $U_{\hat{j}_n}$. We reject H_0 in favor of H_A for

extreme values of the test statistic $\sqrt{n}\hat{\theta}_n$.

2.3. Local behavior of $\hat{\theta}_n$

The challenge of calibrating a test based on $\sqrt{n}\hat{\theta}_n$ is to adapt to its nonregular limiting behavior at $\boldsymbol{\beta}_0 = \mathbf{0}$ (as shown in Theorem 1 below). To accurately capture the asymptotic behavior of $\hat{\theta}_n$ in $\sqrt{n}$-neighborhoods of $\boldsymbol{\beta}_0 = \mathbf{0}$, we consider the local linear model

$$T^{(n)} = \alpha_0 + \boldsymbol{U}^T\boldsymbol{\beta}_n + \varepsilon, \tag{2.6}$$

where $\boldsymbol{\beta}_n = \boldsymbol{\beta}_0 + \boldsymbol{b}_0/\sqrt{n}$, with a local parameter $\boldsymbol{b}_0 \in \mathbb{R}^p$, and ε is unchanged.

Under model (2.6), the observed time and the censoring status are denoted by $X^{(n)} = \min(T^{(n)}, C)$ and $\delta^{(n)} = 1(T^{(n)} \le C)$, respectively. We also define the synthetic response $\tilde{Y}^{(n)}$ and the estimated synthetic response $Y^{(n)}$ in an analogous fashion:

$$\tilde{Y}^{(n)} = \frac{\delta^{(n)}X^{(n)}}{G_0(X^{(n)}-)} \text{ and } Y^{(n)} = \frac{\delta^{(n)}X^{(n)}}{\hat{G}_n(X^{(n)}-)}.$$

For any fixed n, $\tilde{Y}^{(n)}$ has the same mean and covariance with $\boldsymbol{U}$ as those of $T^{(n)}$. The error term associated with $\tilde{Y}^{(n)}$ is $\tilde{\varepsilon}_n = \tilde{Y}^{(n)} - \alpha_0 - \boldsymbol{U}^T\boldsymbol{\beta}_n$, which also has zero mean and is uncorrelated with $\boldsymbol{U}$. Instead of j_0, the label of the predictor most correlated with $T^{(n)}$ is

$$j_n \equiv j(\boldsymbol{\beta}_n) = \arg\max_{j=1,\dots,p} |\mathrm{Corr}(U_j, T^{(n)})| = \arg\max_{j=1,\dots,p} |\mathrm{Corr}(U_j, \tilde{Y}^{(n)})|,$$

and our earlier hypotheses become

$$H_0 : \theta_n = 0 \quad \text{versus} \quad H_A : \theta_n \neq 0,$$

where

$$\theta_n = \frac{\mathrm{Cov}(U_{j_n}, T^{(n)})}{\mathrm{Var}(U_{j_n})}. \tag{2.7}$$

Note that $j_n = j(\boldsymbol{b}_0)$ when $\boldsymbol{\beta}_0 = \mathbf{0}$, but $\boldsymbol{b}_0 \neq \mathbf{0}$, and $j(\boldsymbol{b}_0)$ is assumed unique. Otherwise, j_n is not well defined, and the null hypothesis $\theta_n = 0$ holds when $\boldsymbol{\beta}_0 = \mathbf{0}$ and $\boldsymbol{b}_0 = \mathbf{0}$. If j_0 is unique, then $j_n \to j_0$. The estimators $\hat{j}_n$ and $\hat{\theta}_n$ are now defined by replacing Y by $Y^{(n)}$ in (2.4) and (2.5).

We develop the limiting distribution of $\sqrt{n}\hat{\theta}_n$ in the following theorem under assumptions (A.1)–(A.4) below. The proof is based on the functional delta method (van der Vaart (2000), Chap. 20) and a functional central limit theorem (Pollard (1990), Sec. 10), and is provided in the Supplementary Material.

(A.1) The predictors U_j, for $j = 1, \dots, p$, are bounded, and $|\mathrm{Corr}(U_j, U_k)| < 1$,

for all $j \neq k$.

(A.2) The error term ε in (2.6) has a zero mean and finite variance, and is uncorrelated with $\boldsymbol{U}$.

(A.3) The censoring time C is independent of $(T, \boldsymbol{U})$ and is bounded above by τ (the time to the end of the follow-up).

(A.4) The marginal survival function of the censoring, G_0, is continuous on $\mathcal{T}$, and there exists a positive constant c_g such that $G_0(\tau) > c_g > 0$. In addition, the marginal survival function of T, F_0, is continuous on $\mathcal{T}$, and there exists a positive constant c_f such that $F_0(\tau) > c_f > 0$.

Theorem 1. *Suppose that $j_0 = j(\boldsymbol{\beta}_0)$ is unique when $\boldsymbol{\beta}_0 \neq \mathbf{0}$; $j(\boldsymbol{b}_0)$ is unique when $\boldsymbol{\beta}_0 = \mathbf{0}$ and $\boldsymbol{b}_0 \neq \mathbf{0}$, and that the regularity conditions (A.1)–(A.4) hold. Under the local model* (2.6)*,*

$$\sqrt{n}(\hat{\theta}_n - \theta_n) \xrightarrow{d} \begin{cases} (M_{j_0} + \varphi_{j_0}(\mathbb{L}))/V_{j_0} & \text{if } \boldsymbol{\beta}_0 \neq \mathbf{0}, \\ (M_J + \varphi_J(\mathbb{L}))/V_J + (C_J/V_J - C_{j(\boldsymbol{b}_0)}/V_{j(\boldsymbol{b}_0)})^T \boldsymbol{b}_0 & \text{if } \boldsymbol{\beta}_0 = \mathbf{0}, \end{cases}$$

where $V_j = \operatorname{Var}(U_j)$, $C_j = \operatorname{Cov}(U_j, \boldsymbol{U})$, $J = \arg\max_{j=1,\ldots,p}\{M_j + \varphi_j(\mathbb{L}) + C_j^T \boldsymbol{b}_0\}^2/V_j$, $\boldsymbol{M} = \{M_j, j = 1, \ldots, p\}$ is a mean-zero normal random vector, $\mathbb{L}$ is a mean-zero Gaussian process, and $(\boldsymbol{M}, \mathbb{L})$ is a mean-zero Gaussian process, the covariance of which is provided in the Supplementary Material. The j-indexed functional $\varphi_j \colon \ell_\tau^\infty \to \mathbb{R}$ is defined by

$$\varphi_j(h) = E\left[\frac{(U_j - EU_j)Th(T)}{G_0(T)}\right],$$

where ℓ_τ^∞ denotes the space of bounded functions on $\mathcal{T}$.

Remark 1. The Gaussian process $\mathbb{L}$ is the weak limit of the process $\sqrt{n}(\hat{G}_n - G_0)$. When there is no censoring, $\hat{G}_n(t) = G_0(t) = 1$, for all t, such that $\mathbb{L}$ is a zero process. Then, $\varphi_j(\mathbb{L}) = 0$ for all j, and the limiting distribution reduces to that given by MQ. When there is censoring, $\mathbb{L}$ is a nontrivial Gaussian process and introduces further dispersion into our limiting distribution.

Remark 2. When there is censoring and $\boldsymbol{\beta}_0 \neq \mathbf{0}$, we have T and $\boldsymbol{U}$ correlated, leading to nonzero $\varphi_j(\mathbb{L})$ for all j. Along with the nontrivial process $\mathbb{L}$, the additional term $\varphi_{j_0}(\mathbb{L})$ will be present.

Remark 3. When there is censoring and $\boldsymbol{\beta}_0 = \mathbf{0}$, $\varphi_j(\mathbb{L})$ will vanish everywhere, almost surely (a.s.) for all j, if ε and $\boldsymbol{U}$ are independent. As a result, the additional term $\varphi_J(\mathbb{L})$ disappears. Given the independence between ε and $\boldsymbol{U}$,

the limiting distribution simplifies to

$$\frac{M_J}{V_J} + \left(\frac{C_J}{V_J} - \frac{C_{j(\boldsymbol{b}_0)}}{V_{j(\boldsymbol{b}_0)}}\right)^T \boldsymbol{b}_0.$$

This less complex form of the limiting distribution can be estimated easily from the data. In addition to the possibility of evaluating the asymptotic power (discussed in Section 6), it enables calibration via simulation from the estimated null limiting distribution of $\sqrt{n}\hat{\theta}_n$ (later introduced as "CEND" in Section 5). However, the validity of this approach relies on the highly restrictive assumption that ε and $\boldsymbol{U}$ are independent.

The discontinuity of the limiting distribution at $\boldsymbol{\beta}_0 = \mathbf{0}$ introduces difficulties when designing a screening test based on $\hat{\theta}_n$. If $\boldsymbol{\beta}_0 \neq \mathbf{0}$, naive resampling methods can give consistent estimates of the limiting distribution of $\sqrt{n}(\hat{\theta}_n - \theta_n)$. However, if $\boldsymbol{\beta}_0 = \mathbf{0}$, resampling methods that fail to consider the local behavior of $\sqrt{n}\hat{\theta}_n$ around $\boldsymbol{\beta}_0 = \mathbf{0}$ will give inconsistent estimates of the limiting distribution. To accommodate this nonuniform weak convergence at the point of nonregularity (i.e., $\boldsymbol{\beta}_0 = \mathbf{0}$), our proposed ARTS allows for the flexibility of using different bootstrap strategies to approximate the limiting distribution when $\boldsymbol{\beta}_0 \neq \mathbf{0}$ or $\boldsymbol{\beta}_0 = \mathbf{0}$. Recall that S_j^2 is the sample variance of U_j, for all j. We decompose $\sqrt{n}(\hat{\theta}_n - \theta_n)$ into

$$\sqrt{n}(\hat{\theta}_n - \theta_n)1(|\mathbb{T}_n| > \lambda_n \text{ or } \boldsymbol{\beta}_0 \neq \mathbf{0}) + \sqrt{n}(\hat{\theta}_n - \theta_n)1(|\mathbb{T}_n| \leq \lambda_n, \boldsymbol{\beta}_0 = \mathbf{0}), \quad (2.8)$$

where $\mathbb{T}_n = \sqrt{n}\hat{\theta}_n/\hat{\sigma}_n$ is the maximally selected studentized statistic, and

$$\hat{\sigma}_n^2 = \frac{\mathbb{P}_n(Y - \hat{\alpha}_n - \hat{\theta}_n U_{\hat{j}_n})^2}{S_{\hat{j}_n}^2}$$

with $(\hat{\alpha}_n, \hat{\theta}_n, \hat{j}_n)$ defined in (2.4) and (2.5). The statistic $\mathbb{T}_n$ serves as a pretest to identify the nonregular situation in which we need a more accurate bootstrap strategy to capture the local asymptotic behavior of $\hat{\theta}_n$. Although the asymptotic variance of the KSV estimator in the fixed design case is known (Zhou (1992), Srinivasan and Zhou (1994)), in the present random design case it is simpler to avoid using such a complex standard error estimator. Instead, we base the pretest on the relatively simple statistic $\mathbb{T}_n$. We show that $\hat{\sigma}_n^2$ is asymptotically bounded away from zero and bounded above (the proof is provided in the Supplementary Material). Together with the results in Theorem 1, we prove that $|\mathbb{T}_n| \xrightarrow{a.s.} \infty$ when $\boldsymbol{\beta}_0 \neq \mathbf{0}$, and $|\mathbb{T}_n| = O_p(1)$ when $\boldsymbol{\beta}_0 = \mathbf{0}$. The specification of λ_n is presented in the next section.

We isolate the possibility of $\boldsymbol{\beta}_0 = \mathbf{0}$ by comparing $|\mathbb{T}_n|$ with some screening

threshold λ_n. The first term in (2.8) can be estimated consistently using a centered percentile bootstrap whenever $\lambda_n = o(\sqrt{n})$ and $\lambda_n \to \infty$, because we show $1(|\mathbb{T}_n| > \lambda_n) \xrightarrow{p} 1(\boldsymbol{\beta}_0 \neq \mathbf{0})$ (stated as Lemma 4.1 in the Supplementary Material, along with a detailed proof). Estimating the second term in (2.8) entails additional work. Recall that $\mathbb{P}_n$ is the empirical distribution, P is the distribution of $(X^{(n)}, \delta^{(n)}, \boldsymbol{U})$, and $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$. For $j = 1, \ldots, p$, we define

$$\mathbb{M}_{n,j} = \mathbb{G}_n \tilde{\varepsilon}_n (U_j - \mathbb{P}_n U_j) \text{ and } \mathbb{D}_{n,j} = \sqrt{n}\mathbb{P}_n (U_j - \mathbb{P}_n U_j)(Y^{(n)} - \tilde{Y}^{(n)}).$$

For $\boldsymbol{b} \in \mathbb{R}^p$, we define

$$J_n(\boldsymbol{b}) = \arg \max_{j=1,\ldots,p} \frac{(\mathbb{M}_{n,j} + \mathbb{D}_{n,j} + \mathbb{P}_n (U_j - \mathbb{P}_n U_j) U^T \boldsymbol{b})^2}{S_j^2},$$

and a $\boldsymbol{b}$-indexed process

$$\mathbb{Q}_n(\boldsymbol{b}) = \frac{(\mathbb{M}_{n,J_n(\boldsymbol{b})} + \mathbb{D}_{n,J_n(\boldsymbol{b})} + \mathbb{P}_n (U_{J_n(\boldsymbol{b})} - \mathbb{P}_n U_{J_n(\boldsymbol{b})}) U^T \boldsymbol{b})}{S^2_{J_n(\boldsymbol{b})}} - \frac{C^T_{j(\boldsymbol{b})} \boldsymbol{b}}{V_{j(\boldsymbol{b})}}.$$

Below, we express the second term in (2.8) as a function $\mathbb{Q}_n(\boldsymbol{b}_0)$. When $\boldsymbol{\beta}_0 = \mathbf{0}$, it is easy to see that

$$\begin{aligned}
\sqrt{n}\hat{\theta}_j &= \frac{\sqrt{n}\mathbb{P}_n (U_j - \mathbb{P}_n U_j)\tilde{Y}^{(n)}}{S_j^2} + \frac{\sqrt{n}\mathbb{P}_n (U_j - \mathbb{P}_n U_j)(Y^{(n)} - \tilde{Y}^{(n)})}{S_j^2} \\
&= \frac{(\mathbb{G}_n \tilde{\varepsilon}_n (U_j - \mathbb{P}_n U_j) + \sqrt{n}\mathbb{P}_n (U_j - \mathbb{P}_n U_j)(Y^{(n)} - \tilde{Y}^{(n)}) + \mathbb{P}_n (U_j - \mathbb{P}_n U_j) \boldsymbol{U}^T \boldsymbol{b}_0)}{S_j^2} \\
&= \frac{(\mathbb{M}_{n,j} + \mathbb{D}_{n,j} + \mathbb{P}_n (U_j - \mathbb{P}_n U_j) U^T \boldsymbol{b}_0)}{S_j^2},
\end{aligned}$$

for all j. Along with $\hat{j}_n = J_n(\boldsymbol{b}_0)$ and $j_n = j(\boldsymbol{b}_0)$ when $\boldsymbol{\beta}_0 = \mathbf{0}$, we have $\sqrt{n}\theta_n = C^T_{j(\boldsymbol{b}_0)} \boldsymbol{b}_0 / V_{j(\boldsymbol{b}_0)}$ and therefore, $\sqrt{n}(\hat{\theta}_n - \theta_n) = \mathbb{Q}_n(\boldsymbol{b}_0)$. Hence, the decomposition of $\sqrt{n}(\hat{\theta}_n - \theta_n)$ can be expressed as

$$\sqrt{n}(\hat{\theta}_n - \theta_n) = \sqrt{n}(\hat{\theta}_n - \theta_n) 1(|\mathbb{T}_n| > \lambda_n \text{ or } \boldsymbol{\beta}_0 \neq \mathbf{0}) + \mathbb{Q}_n(\boldsymbol{b}_0) 1(|\mathbb{T}_n| \leq \lambda_n, \boldsymbol{\beta}_0 = \mathbf{0}). \tag{2.9}$$

In Theorem 2 below, we show that $\mathbb{Q}_n(\boldsymbol{b})$ can be consistently bootstrapped for any given $\boldsymbol{b}$. Provided that $\boldsymbol{b}_0$ is known, we can directly bootstrap the expression in (2.9) to consistently estimate the limiting distribution of $\sqrt{n}(\hat{\theta}_n - \theta_n)$. Hereafter, the superscript $*$ is used to indicate the bootstrap version of an estimator.

Theorem 2. *Suppose that all conditions for Theorem 1 hold, and the tuning parameter λ_n satisfies $\lambda_n = o(\sqrt{n})$ and $\lambda_n \to \infty$ as $n \to \infty$. Under the local model* (2.6)*,*

$$\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)1(|\mathbb{T}_n^*| > \lambda_n \text{ or } |\mathbb{T}_n| > \lambda_n) + \mathbb{Q}_n^*(\boldsymbol{b}_0)1(|\mathbb{T}_n^*| \le \lambda_n, |\mathbb{T}_n| \le \lambda_n)$$

converges to the limiting distribution of $\sqrt{n}(\hat{\theta}_n - \theta_n)$ *conditionally (on the data) in probability.*

2.4. ARTS screening procedure

The ARTS screening procedure uses a bootstrap calibration for the test statistic $\sqrt{n}\hat{\theta}_n$ based on a special case of Theorem 2, specifically, $\boldsymbol{b}_0 = \mathbf{0}$. To approximate the limiting distribution of $\sqrt{n}\hat{\theta}_n$ under the null, it suffices to bootstrap

$$B_n = \sqrt{n}(\hat{\theta}_n - \theta_n)1(|\mathbb{T}_n| > \lambda_n \text{ or } \boldsymbol{\beta}_0 \neq \mathbf{0}) + \mathbb{Q}_n(\mathbf{0})1(|\mathbb{T}_n| \le \lambda_n, \boldsymbol{\beta}_0 = \mathbf{0}), \quad (2.10)$$

and the corresponding bootstrap version is

$$B_n^* = \sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)1(|\mathbb{T}_n^*| > \lambda_n \text{ or } |\mathbb{T}_n| > \lambda_n) + \mathbb{Q}_n^*(\mathbf{0})1(|\mathbb{T}_n^*| \le \lambda_n, |\mathbb{T}_n| \le \lambda_n). \quad (2.11)$$

For some nominal level α, define the critical values c_l and c_u, respectively, by the lower and upper $100(\alpha/2)$-th percentiles of 1,000 replications of B_n^*. We reject the null hypothesis, and conclude that there is at least one significant predictor if $\sqrt{n}\hat{\theta}_n$ falls outside the interval $[c_l, c_u]$.

Given the conditions that $\lambda_n = o(\sqrt{n})$ and $\lambda_n \to \infty$, the pretest demonstrates an asymptotically negligible Type-I error rate $P(|\mathbb{T}_n| > \lambda_n | \theta_n = 0) \to 0$, because we have shown that $P(|\mathbb{T}_n| > \lambda_n) \to 1(\boldsymbol{\beta}_0 \neq \mathbf{0})$ in Lemma 4.1, stated in the Supplementary Material. Provided that $\tilde{\varepsilon}$ and $\boldsymbol{U}$ are independent, a special case of Theorem 1 indicates that $\mathbb{T}_n \xrightarrow{d} \max_{j=1,\ldots,p} |Z_j|$ at the null, where $\{Z_j, j = 1, \ldots, p\}$ is a vector of standard normal random variables. Using similar arguments to those of MQ, the asymptotic Type-I error rate of the pretest can be controlled below level α if we set $\lambda_n \ge \Phi^{-1}(1 - \alpha/(2p))$, where Φ denotes the standard normal distribution function. To satisfy the conditions that $\lambda_n = o(\sqrt{n})$ and $\lambda_n \to \infty$, one reasonable selection of the threshold would be $\lambda_n = \max\{\sqrt{a \log n}, \Phi^{-1}(1 - \alpha/(2p))\}$, for some constant $a > 0$.

To determine the value of the constant a in practice, we use a double-bootstrap. That is, we produce 1,000 bootstrap estimates $\hat{\theta}_n^*$, and apply the ARTS to a further 1,000 nested double-bootstrap samples to obtain the acceptance region $[c_l^*, c_u^*]$ for each $\hat{\theta}_n^*$. If the test statistic $\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)$ falls outside $[c_l^*, c_u^*]$, we record this as a rejection. The constant a is specified as the value that results in 5% of these 1,000 ARTS procedures being rejected. This data-driven selection of a is adopted in our numerical studies and applications to real data.

Note that in each bootstrap and nested double-bootstrap sample, we set τ as the 90% empirical percentile of the observed time and control the censoring rate around the same level, as in the original data.

3. ARTS Adjusted for Baseline Covariates

When screening high-dimensional predictors of survival outcomes, it is common practice to adjust for baseline demographic and clinical covariates. These baseline covariates include age, disease stage, tumor thickness, and lymph node status; in the DLBCL study, we have the International Prognostic Index (IPI). The IPI is a widely used prognostic index that reflects the combination of clinical covariates (cf., The International Non-Hodgkin's Lymphoma Prognostic Factors Project (1993)). Such baseline covariates (with moderate dimensionality) do not need to be screened, but do need to be incorporated as covariates in the AFT model. In this section, we modify the ARTS (as *adjusted ARTS*) to account for the effect of these covariates.

Let $\tilde{\boldsymbol{U}} = (\tilde{U}_1, \ldots, \tilde{U}_q)^T$ be a vector of baseline covariates. With $\tilde{\boldsymbol{U}}$ included, the true AFT model (1.1) can be expressed as

$$T = \alpha_0 + \boldsymbol{U}^T\boldsymbol{\beta}_0 + \tilde{\boldsymbol{U}}^T\boldsymbol{\gamma}_0 + \varepsilon, \tag{3.1}$$

where $\boldsymbol{\gamma}_0 \in \mathbb{R}^q$, $\tilde{\boldsymbol{U}}$ is assumed to be bounded, and the error term ε is uncorrelated with $\tilde{\boldsymbol{U}}$. We wish to test whether $\boldsymbol{\beta}_0 = \mathbf{0}$, which includes an adjustment for $\tilde{\boldsymbol{U}}$. Projecting $\tilde{\boldsymbol{U}}$ on the space spanned by $\boldsymbol{U}$, we reformulate the AFT model (3.1) as

$$T = \alpha_0' + \boldsymbol{D}^T\boldsymbol{\beta}_0 + \varepsilon', \tag{3.2}$$

where $\boldsymbol{D} = (D_1, \ldots, D_p)^T$ with $D_j = U_j - \tilde{\alpha}_j - \tilde{\boldsymbol{U}}^T\tilde{\boldsymbol{\gamma}}_j$; at the same time,

$$\begin{aligned}(\tilde{\alpha}_j, \tilde{\boldsymbol{\gamma}}_j^T) &= \left(E[U_j] - E[\tilde{\boldsymbol{U}}^T\tilde{\boldsymbol{\gamma}}_j], (\Sigma_{\tilde{\boldsymbol{U}}}^{-1}\mathrm{Cov}(U_j, \tilde{\boldsymbol{U}}))^T\right),\\ \alpha_0' &= \alpha_0 + (\tilde{\alpha}_1, \ldots, \tilde{\alpha}_p)\boldsymbol{\beta}_0 + E[\tilde{\boldsymbol{U}}^T((\tilde{\gamma}_1, \ldots, \tilde{\gamma}_p)\boldsymbol{\beta}_0 + \boldsymbol{\gamma}_0)],\\ \varepsilon' &= \tilde{\boldsymbol{U}}^T((\tilde{\gamma}_1, \ldots, \tilde{\gamma}_p)\boldsymbol{\beta}_0 + \boldsymbol{\gamma}_0) - E[\tilde{\boldsymbol{U}}^T((\tilde{\gamma}_1, \ldots, \tilde{\gamma}_p)\boldsymbol{\beta}_0 + \boldsymbol{\gamma}_0)] + \varepsilon,\end{aligned}$$

and $\Sigma_{\tilde{\boldsymbol{U}}}$ is the covariance matrix of $\tilde{\boldsymbol{U}}$. Note that $\tilde{\alpha}_j + \tilde{\boldsymbol{U}}^T\tilde{\boldsymbol{\gamma}}_j$ is the best linear unbiased predictor of U_j based on $\tilde{\boldsymbol{U}}$. According to the definition of $(\tilde{\alpha}_j, \tilde{\boldsymbol{\gamma}}_j)$, it is obvious that $E[D_j] = 0$ and $\mathrm{Cov}(D_j, \tilde{\boldsymbol{U}}^T\boldsymbol{\gamma}) = 0$, for all j and any vector $\boldsymbol{\gamma} \in \mathbb{R}^q$. The new error term ε' inherits the properties of ε and satisfies the moment conditions required for the ARTS: $E[\varepsilon'] = 0$, $E[(\varepsilon')^2] < \infty$, and ε' is uncorrelated with $\boldsymbol{D}$. To test whether $\boldsymbol{\beta}_0 = \mathbf{0}$ under model (3.2), it suffices to

test

$$H_0 : \theta_0' = 0 \quad \text{versus} \quad H_A : \theta_0' \neq 0,$$

where $\theta_0' = \mathrm{Cov}(D_{j'(\boldsymbol{\beta}_0)}, T)/\mathrm{Var}(D_{j'(\boldsymbol{\beta}_0)})$, and $j'(\boldsymbol{b}) = \arg\ \max_{j=1,\ldots,p} |\mathrm{Corr}(D_j, \boldsymbol{D}^T\boldsymbol{b})|$ for any $\boldsymbol{b} \in \mathbb{R}^p$, implying $j'(\boldsymbol{\beta}_0) = \arg\ \max_{j=1,\ldots,p} |\mathrm{Corr}(D_j, T)|$.

The adjusted ARTS regresses each screening predictor on baseline covariates and applies the ARTS with the corresponding residuals $\hat{\boldsymbol{D}} = (\hat{D}_1, \ldots, \hat{D}_p)^T$ as predictors. Because $\hat{D}_j$ involves a least-squares-type estimate of $(\tilde{\alpha}_j, \tilde{\boldsymbol{\gamma}}_j)$ for $j = 1, \ldots, p$, we can use the strong consistency of the estimates over all j (implied by SLLN and fixed p) to justify the replacement of $\boldsymbol{D}$ by $\hat{\boldsymbol{D}}$. The bootstrap consistency is also guaranteed. Thus, we only need to resample the residuals in the procedures of the bootstrap and double-bootstrap. This offers a considerable saving in terms of computation cost (caused by implementing projections every time we have bootstrap or double-bootstrap samples), especially when p is large. We tailor the adjustment of $\tilde{\boldsymbol{U}}$ to fit within the ARTS framework to avoid using a test statistic in matrix form, which is inevitable when fitting a multi-variable AFT model to adjust for $\tilde{\boldsymbol{U}}$. This idea is crucial because it has the advantage of extending the theoretical results developed for the ARTS to the adjusted ARTS.

4. Forward-stepwise ARTS

Given one significant predictor detected by the ARTS, it is natural to continue searching for other potential predictors, conditional on the information provided by the identified predictor. We implement the idea used in the adjusted ARTS procedure to fulfill this task in a forward and stepwise direction. The conditional screening continues until no further significance is detected. We refer to this screening procedure as the *forward-stepwise ARTS*, implemented as follows:

1. Given the predictor $U_{\hat{j}_n}$ detected by the ARTS, obtain residuals from regressing U_j on $U_{\hat{j}_n}$ whenever $j \neq \hat{j}_n$. Treat the residuals as screened predictors and run the adjusted ARTS. If no significant results are returned, stop the procedure; otherwise, collect the newly found significant predictor $U_{\tilde{j}_n}$.
2. Use the residuals from regressing U_j on $(U_{\hat{j}_n}, U_{\tilde{j}_n})$ as updated predictors, for all $j \notin (\hat{j}_n, \tilde{j}_n)$. Implement the adjusted ARTS based on these updated predictors, in order to detect the next significant predictor.
3. Keep accumulating predictors until no further significant predictors are detected.

Our forward-stepwise ARTS procedure successively updates the predictors using the residuals from regressing on previously identified predictors. Compared with the residual analysis suggested by MQ, our forward-stepwise procedure allows the regression coefficients of all already included predictors to be refitted at each step. This implies the detection of further significant predictors, adjusting for those already-included.

5. Competing Methods

We compare the performance of the ARTS with several procedures that are widely applied to detect the presence of significant predictors for the survival outcome. When considering the adjustment of baseline covariates, these procedures can be modified as alternatives to the adjusted ARTS procedure.

5.1. AFT model approaches

Marginal parametric AFT models with Bonferroni correction (BONF-AFT). A marginal parametric AFT model is often used to predict T from each predictor by specifying a parametric form of the error distribution, from which we obtain the maximum likelihood estimate of the marginal regression coefficient of each predictor. A Z-test with a Bonferroni correction is carried out to test whether each marginal regression coefficient is zero. This method can be implemented using the `survreg` function from the `survival` package of R. To adjust for baseline covariates, we treat the residual $\hat{D}_j$ as the predictor in a marginal parametric AFT model, for $j = 1, \ldots, p$. In our finite-sample simulations, we specify that the error term follows a standard normal distribution.

Marginal AFT models with higher criticism correction (HC). The higher criticism method is a test proposed by John Tukey for determining the overall significance of a collection of independent p-values. We use the statistic developed by Donoho and Jin, which is expected to perform well if the predictors are nearly uncorrelated (Donoho and Jin (2004), Donoho and Jin (2015)).

Centered percentile bootstrap with AFT model (CPB-AFT). In contrast to the ARTS, this procedure works on the premise that there is at least one active predictor. Thus, it only bootstraps the first part of (2.9) to estimate the upper and lower $100(\alpha/2)$-th percentiles of the limiting distribution of $\sqrt{n}(\hat{\theta}_n - \theta_n)$. The estimated percentiles provide critical values for the test statistic $\sqrt{n}\hat{\theta}_n$ (Efron and Tibshirani (1993)). Note that this method yields a special case of the ARTS with $\lambda_n = 0$. We can easily modify this method to adjust for baseline covariates by

replacing θ_n and $\hat{\theta}_n$ with their counterparts in the framework given in Section 3.

Calibration by simulation from the estimated null distribution (CEND). The asymptotic acceptance region is used to calibrate the test, and can be constructed from the special case in which ε and $\boldsymbol{U}$ are independent. Here, we simulate the limiting distribution of the scaled test statistic $\sqrt{n}\hat{\theta}_n/s$ under the null, where $s^2 = \mathbb{P}_n(Y_i^{(n)} - \hat{\alpha}_n - \hat{\theta}_n U_{\hat{j}_n})^2$. At the null, Theorem 1 implies that $\sqrt{n}\hat{\theta}_n/s \xrightarrow{d} \tilde{M}_J/V_J$, where $\{\tilde{M}_j, j = 1, \ldots, p\} \sim \mathcal{N}_p(\mathbf{0}, \Sigma_{\boldsymbol{U}})$, $\Sigma_{\boldsymbol{U}}$ is the covariance matrix of $\boldsymbol{U}$, and $J = \arg\max_j \tilde{M}_j^2/V_j$. With $\Sigma_{\boldsymbol{U}}$ estimated using the sample covariance matrix of $\boldsymbol{U}$, we generate 1,000 realizations from $\mathcal{N}_p(\mathbf{0}, \Sigma_{\boldsymbol{U}})$, which we to obtain 1,000 random copies of $\sqrt{n}\hat{\theta}_n$. Then, we use the corresponding percentiles to develop the acceptance region. We reject the null hypothesis if $\sqrt{n}\hat{\theta}_n$ falls outside this region. The version that adjusts for baseline covariates can be developed analogously by taking $\hat{D}$ as predictors.

5.2. Cox model approaches

The other popular approach for linking predictors to the survival outcome is the Cox model, where the related statistical inference can be developed based on the partial likelihood (Cox (1972), Cox (1975)).

Partial likelihood ratio test (PLRT). This test uses the likelihood ratio test statistic Λ, which is the ratio of the partial likelihood from the full Cox model to that from the reduced model at the null. Provided that $\Lambda \xrightarrow{d} \chi_p^2$ (chi-square distribution with p degrees of freedom), comparing Λ with a χ_p^2-distributed random variable gives the p-value to calibrate the test. However, the PLRT is only feasible in the case of $n > p$, because it involves a full linear model containing all of the predictors. To adjust for baseline covariates, we define the test statistic as the ratio of the partial likelihood from a Cox model containing $(\boldsymbol{U}, \tilde{\boldsymbol{U}})$ to that from a Cox model considering $\tilde{\boldsymbol{U}}$ only. This statistic weakly converges to χ_p^2.

Marginal Cox models with Bonferroni correction (BONF-COX). This procedure is similar to the BONF-AFT, but is based on marginal Cox models to link the survival outcome to each predictor U_j, for $j = 1, \ldots, p$. Given the asymptotic normality of the maximum partial likelihood estimator (MPLE) (Andersen and Gill (1982)), we conduct a Z-test with a Bonferroni correction to investigate whether each marginal regression coefficient is zero. To adjust for baseline covariates, we can instead fit Cox models containing $(U_j, \tilde{\boldsymbol{U}})$ for all j, and use the corresponding MPLE of the regression coefficient of U_j as the test statistic.

Centered percentile bootstrap with Cox model (CPB-COX). This procedure is similar to the CPB-AFT in general, but the selected predictor is determined in a different fashion. The marginal p-values are obtained from Z-tests based on separate marginal Cox models, and we select the predictor that marginally introduces the minimal p-value. We apply a centered percentile bootstrap on the MPLE of the regression coefficient of this selected predictor (i.e., the most significant predictor). To consider additional baseline covariates, we consider Cox models containing $(U_j, \tilde{\boldsymbol{U}})$, for all j, and bootstrap the MPLE of the regression coefficient of the most significant predictor among U_j, while adjusting for $\tilde{\boldsymbol{U}}$.

Global test based on Cox model (GLOBAL). A score test is proposed to investigate whether the predictors $\boldsymbol{U}$ contribute to the hazard rate (Goeman et al. (2005)). The components of $\boldsymbol{\beta}_0$ are assumed to be random and independently follow a prior distribution with mean zero and common variance v. Here, it suffices to test whether $v = 0$ to investigate whether $\boldsymbol{\beta}_0 = \mathbf{0}$. Let $\mathbf{r} = (r_1, \ldots, r_n)^T$, with $r_i = \boldsymbol{U}_i^T \boldsymbol{\beta}_0$ for all i, and note that $\mathbf{r}$ is not observed because the unknown parameter vector $\boldsymbol{\beta}_0$ is included. By the assumptions on $\boldsymbol{\beta}_0$, $\mathbf{r}$ has mean zero and covariance matrix $v\boldsymbol{U}\boldsymbol{U}^T$. Under the noninformative censoring assumption, the marginal likelihood function of v is defined by

$$L(v) = E_{\mathbf{r}}\left[\exp\left(\sum_{i=1}^{n}[\delta_i(\ln(h_0(X_i)) + r_i) - \exp(r_i)H_0(X_i)]\right)\right], \tag{5.1}$$

where $H_0(t) = \int_0^t h_0(s)ds$ is the cumulative baseline hazard function up to time t. Applying the second-order Taylor expansion to the exponential term in (5.1) with respect to $\mathbf{r}$, $L(v)$ can be expressed by the first and second moments of $\mathbf{r}$ (Le Cessie and van Houwelingen (1995)). This implies that we can establish the desired test statistic in terms of the score function of v, which only involves the first and second moments of $\boldsymbol{\beta}_0$, without specifying the prior distribution. There are two ways to calculate the p-value: using asymptotic theory, and using permutation arguments. We compare both to the ARTS in our numerical studies. This global test can be modified to adjust for baseline covariates by simultaneously including $\boldsymbol{U}$ and $\tilde{\boldsymbol{U}}$ in the Cox model, and the test statistic is constructed conditional on the MPLE of the regression coefficients of $\tilde{\boldsymbol{U}}$.

6. Numerical Studies

6.1. Finite-sample simulations

The performance of the ARTS is evaluated using numerical studies under

different data-generating scenarios. The underlying survival outcome can follow either an AFT model or a proportional hazards model. For the former, we consider three data-generating models:

Model 1 $T = \varepsilon$;
Model 2 $T = U_1/4 + \varepsilon$;
Model 3 $T = \sum_{j=1}^{p} \beta_j U_j + \varepsilon$ with $\beta_1 = \cdots = \beta_5 = 0.15$, $\beta_6 = \cdots = \beta_{10} = -0.1$, and $\beta_j = 0$ for $j \geq 11$,

where ε denotes the noise, which follows a standard normal distribution and is independent of $\boldsymbol{U}$. In Model 1, there is no active predictor, whereas there is only a single active predictor in Model 2. In Model 3, we have 10 active predictors and the most correlated predictor is not unique. The censoring time C follows an exponential distribution with various rate parameters for light censoring (10% of subjects with censored survival outcomes), moderate censoring (20%), and heavy censoring (40%). The vector of predictors $\boldsymbol{U}$ follows a p-dimensional normal distribution with each component $U_j \sim \mathcal{N}(0, 1)$, and an exchangeable correlation structure $\mathrm{Corr}(U_j, U_k) = 0.5$ for $j \neq k$.

We also generate the survival outcome based on the following proportional hazards models (Bender, Austin and Blettner (2005)):

Model 4 $h(t|\boldsymbol{U}) = 2\exp(t)$;
Model 5 $h(t|\boldsymbol{U}) = 2\exp(t)\exp(U_1/4)$;
Model 6 $h(t|\boldsymbol{U}) = 2\exp(t)\exp(\sum_{j=1}^{p} \beta_j U_j)$ with the value of $(\beta_1, \ldots, \beta_p)$ as stated in **Model 3**.

To achieve the designed censoring rates, we generate the censoring time as an exponential random variable, for various choices of the rate parameter. We use Models 1 and 4 as the null models, Models 2 and 5 as the alternative models with a sparse signal, and Models 3 and 6 as the alternative models with weak dense signals.

For each data-generating scenario, we consider two sample sizes ($n = 100$ and 200), and five values for the dimension of the predictors ($p = 10$, 50, 100, 150, and 200). A nominal significance level of 5% is used throughout. The number of bootstrap replications is set as 1,000. The selection of the threshold λ_n follows the steps stated in Section 2.4. To provide a full comparison, we compare the performance of the ARTS with the competing methods introduced in Section 5. The empirical rejection rates based on 1,000 Monte Carlo replications under various censoring rates are displayed in Figures 1–2. The panels for Models 1

and 4 give Type-I error rates, which we compare using the nominal level of 5%. The panels for Models 2–6 indicate the power of each test.

In Figure 1, the ARTS controls the Type-I error rates (or equivalently, FWERs) around the nominal level, and demonstrates relatively high power for all alternative models. The BONF-AFT method gives more conservative Type- I error rates and lower power than the ARTS, with the exception of achieving similar power to the ARTS under alternative models with heavy censoring and $n = 200$. The HC method is anti-conservative and fails to control the Type-I errors. We suspect this is due to the relatively high correlation between the predictors, for which HC is not designed. The BONF-COX method and the global test based on asymptotic theory (GLOBAL-asymp) are highly conservative and lead to low power. Both the CPB-AFT and the CPB-COX are anti-conservative, with the empirical Type-I error rates considerably exceeding the nominal level under different sample sizes and various censoring rates (and thus going out of range somewhere in the left panels of Figure 1). The global test based on the permutation arguments (GLOBAL-permut) takes good control of the Type-I error rates, but claims much lower power than the ARTS, especially under light or moderate censoring. Both the CEND and the PLRT exhibit poor performance: the former yields large Type-I error rates but low power, whereas the latter introduces extremely high Type-I error rates. (The results of the PLRT are not shown here.) The unsatisfying performance of the CEND may result from small sample sizes in the simulations, given that the CEND is developed based on a simplified form of the limiting distribution. The power of each approach rises as the sample size increases and the censoring rate decreases. A comparison between the results of Models 2 and 3 shows no adverse impact on the power of the ARTS when the maximally correlated predictor is nonunique.

In Figure 2, where the data are not generated from AFT models, the ARTS retains good control of the Type-I error rates. On the other hand, the power of the ARTS is unstable when $n = 100$ or in the case of heavy censoring. Under light or moderate censoring, the power of the ARTS under Models 5 and 6 deteriorates sharply when $n = 100$ and p increases, whereas the ARTS maintains stable power when $n = 200$. With a misspecified error distribution, the BONF-AFT surprisingly controls the Type-I error rates well, but leads to much worse power. In contrast, the BONF-COX yields relatively greater power when the underlying survival outcome is generated from the proportional hazards model, although it is still conservative at the null. Other competing methods present similar results to those in Figure 1. Despite being unstable in terms of power owing to model

misspecification, the ARTS still strikes a better balance between controlling the Type-I error and achieving sufficient power than other methods do, especially for light or moderate censoring and a large sample size. Comparing Figure 1 with Figure 2, we find that the ARTS is less susceptible to model misspecification than competing methods are. In the scenarios of the AFT data-generating models, the ARTS apparently dominates the Cox model approaches throughout; in the scenarios where the data are generated from proportional hazards models, the ARTS still exhibits better performance in the FWER and power than that of the Cox-model-relevant approaches when the censoring is light or moderate and $n = 200$.

6.2. Screening performance of ARTS

We further assess the performance of the ARTS as a full screening method (i.e., retaining all covariates with marginal test statistics beyond the critical values calculated for $\sqrt{n}\hat{\theta}_n$) in terms of the false discovery rate (FDR), false negative rate (FNR), and false positive rate (FPR). Using a simulation study, we compare the screening performance of the ARTS with the Benjamini–Hochberg procedure (BH, Benjamini and Hochberg (1995)) and the Holm–Bonferroni procedure (HB, Holm (1979)). Relevant results are presented in Section S5 of the Supplementary Material.

The power (as given by the average values of (1 - FNR)) is slightly less for the ARTS than for the BH, which is expected because the acceptance region is constructed from the critical values of the maximum correlation statistic $\hat{\theta}_n$, leading to results that are more conservative. We expect, however, that the forward-stepwise ARTS will outperform the ARTS screening procedure because it re-calibrates at each step. In terms of the FDR and FPR, the performance of the ARTS and BH are comparable, although that of the Bonferroni method is more conservative as expected. The HB and Bonferroni methods show similar performance with respect to all the measures.

6.3. Asymptotic power evaluation

In this section, we conduct a simulation study to evaluate the asymptotic FWER and the power of the ARTS, as compared with those of the BONF-AFT. We assess the asymptotic FWER and power based on the limiting distribution shown in Theorem 1. This approach can be a computationally efficient alternative to the simulation method used in our finite-sample studies, because it avoids the required double-bootstrap (for threshold selection) that incurs a heavy com-

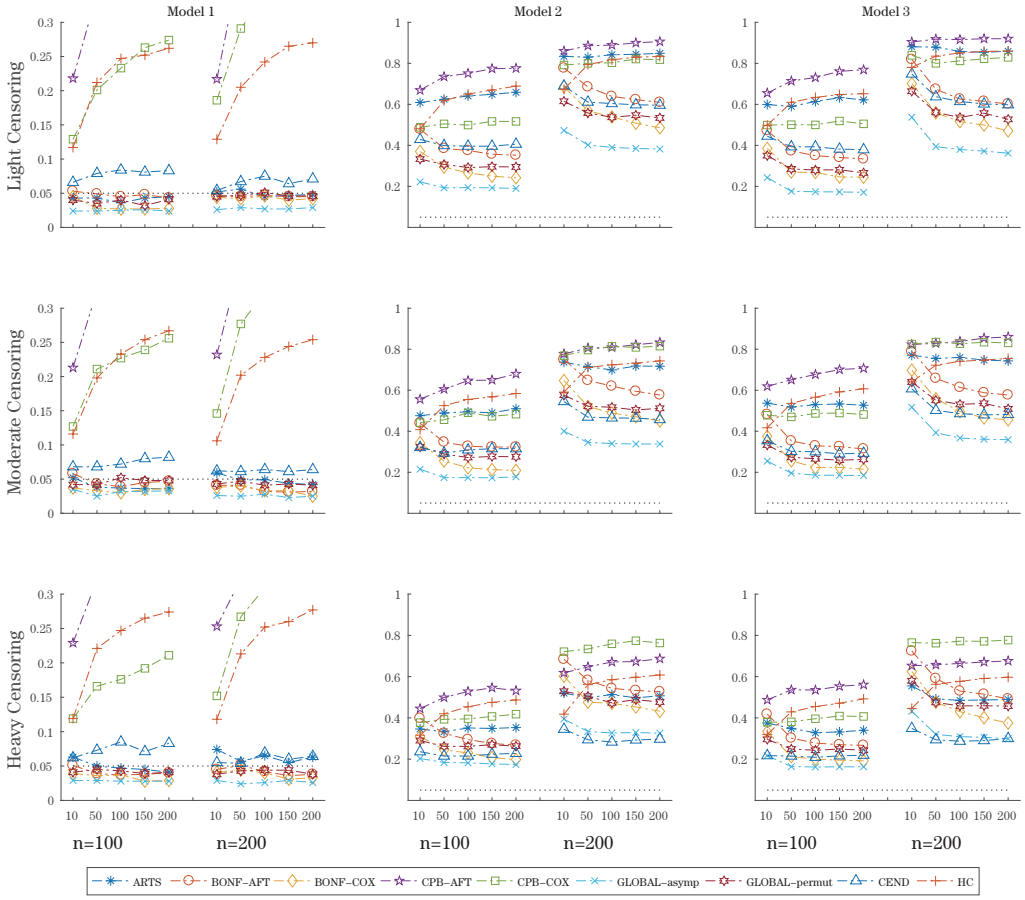

Figure 1. Empirical rejection rates based on 1,000 samples generated from Models 1–3, with the dimension ranging from $p = 10$ to $p = 200$.

putation when implementing the ARTS.

Owing to the complicated limiting distribution shown in Theorem 1, this approach is only feasible when $\varphi_j(\mathbb{L})$ can be reasonably negligible for all j. One possible situation is when $\boldsymbol{\beta}_0 = \mathbf{0}$ and the error term ε is independent of $\boldsymbol{U}$. This restriction on ε facilitates the evaluation of the asymptotic FWER at the null ($\boldsymbol{\beta}_0 = \mathbf{0}$, $\boldsymbol{b}_0 = \mathbf{0}$) and the asymptotic power at local alternatives ($\boldsymbol{\beta}_0 = \mathbf{0}$, $\boldsymbol{b}_0 \neq \mathbf{0}$). This offers a saving in terms of computational costs, at the price of being sensitive to model misspecification.

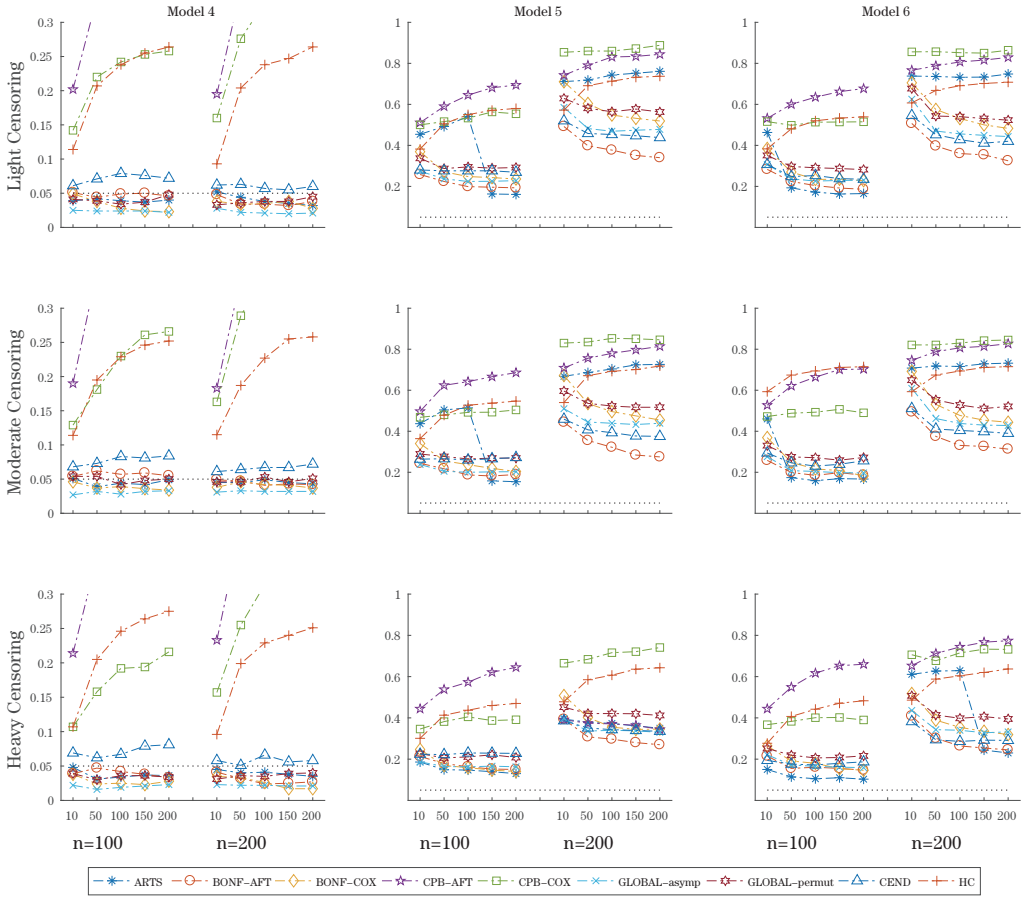

Figure 2. Empirical rejection rates based on 1,000 samples generated from Models 4–6, with the dimension ranging from $p = 10$ to $p = 200$.

Consider a local model

$$T^{(n)} = (n^{-1/2}b_0)U_1 + \varepsilon, \tag{6.1}$$

where U_1 is the first element of $\boldsymbol{U}$. The predictors $\boldsymbol{U}$, the error term ε, and the censoring time C are generated as in Section 6.1. We allow b_0 to vary over a grid in [0, 5] by increments of 0.5. Under this local model, the complex limiting distribution reduces to a simpler form:

$$\sqrt{n}(\hat{\theta}_n - \theta_n) \xrightarrow{d} \frac{(M_J + b_0\mathrm{Cov}(U_J, U_1))}{\mathrm{Var}(U_J)} - b_0, \tag{6.2}$$

where $J = \arg\max_j\{M_j + b_0\mathrm{Cov}(U_j, U_1)\}^2/\mathrm{Var}(U_j)$, and $\boldsymbol{M} = \{M_j, j = 1, \ldots, p\}$ is a mean-zero normal random vector with a covariance matrix given by that of

the random vector $\{\tilde{\varepsilon}(U_j - EU_j), j = 1, \ldots, p\}$. This evaluation procedure is implemented as follows.

1. For each value of b_0 on the grid, generate a large sample (with $n = 10,000$) from the local model (6.1) and compute the corresponding $Y^{(n)}$. Using a fixed threshold λ_n, use the ARTS to develop the acceptance region $[c_l, c_u]$ based on this sample.

2. For each given b_0, take 10,000 draws from the limiting distribution in (6.2), and then obtain 10,000 realizations of $\sqrt{n}\hat{\theta}_n$.

3. The asymptotic rejection rate of the ARTS (for the given b_0) is assessed by computing the proportion of the realizations that fall outside $[c_l, c_u]$ from the 10,000 realizations of $\sqrt{n}\hat{\theta}_n$.

To reflect the random variation of the asymptotic FWER and the power over the samples generated in Step 1, we independently implement the above procedure 20 times and display the corresponding asymptotic rejection rates in a box plot for each b_0. For comparison, we also plot the asymptotic power of the BONF-AFT, which is approximated by the rejection rate from 1,000 samples, each of size $n = 10{,}000$.

To make the above evaluation practical for large p, say $p = 1{,}000$, the threshold λ_n is fixed at $0, 4.3, 6.1$, and 7.4 as the constant a takes corresponding values of $0, 2, 4$, and 6. We present the results under light censoring (Figure 3), moderate censoring (Figure 4), and heavy censoring (Figure 5). Because the plots are similar between $a = 0$ and $a = 1$ and have no obvious difference when $a \geq 6$, we only present the results for $a = 0, 2, 4, 6$, for conciseness. From these figures, we observe that smaller values of a lead to the ARTS yielding results that are more anti-conservative, as observed in previous numerical studies. When $a = 0$, in particular, the ARTS reduces to the CPB-AFT. On the other hand, the ARTS behaves more stably and provides more accurate control of the Type-I error rates as a increases. In addition, the variation within each box plot decreases when the value of a increases.

Comparing the asymptotic power of the BONF-AFT (denoted by the circle) with the median of each box plot, we find that the ARTS has more satisfactory performance than that of the BONF-AFT in most cases. In terms of median power, the ARTS even provides an extra 20% power in some situations (e.g., at $b_0 = 3$, when $a = 4$ or $a = 6$ for all types of censoring). To control the asymptotic FWER, a reasonable choice is $a = 4$ under light or moderate censoring, because

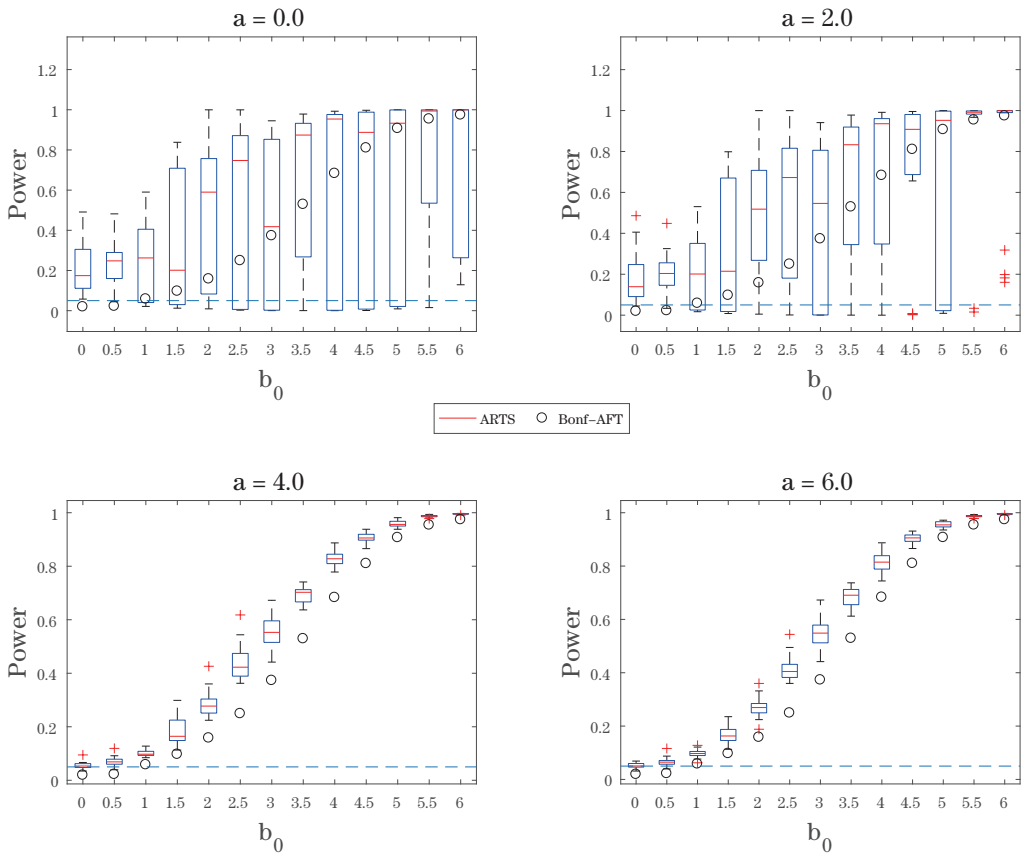

Figure 3. Asymptotic Type-I error and power of ARTS compared with BONF-AFT for $p = 1{,}000$ under light censoring, where ARTS is implemented with a fixed threshold λ_n specified by $a = \{0, 2, 4, 6\}$, and each box plot is based on 20 independent replications with $n = 10,000$.

the median FWER starts to touch the nominal level and the corresponding variation within the box plot diminishes. On the other hand, the selection of a should fall between 2 and 4 under heavy censoring, because the median FWER remains higher than 5% when $a = 2$, but drops below 5% at $a = 4$.

6.4. Error dependent on predictors

In this section, we present the control on the FWER of the ARTS, when the error term ε is still uncorrelated with but dependent on the predictors $\boldsymbol{U}$. For simplicity, $\boldsymbol{U}$ follows a p-dimensional normal distribution with mean zero and an identity covariance matrix, implying that the predictors are independent of each

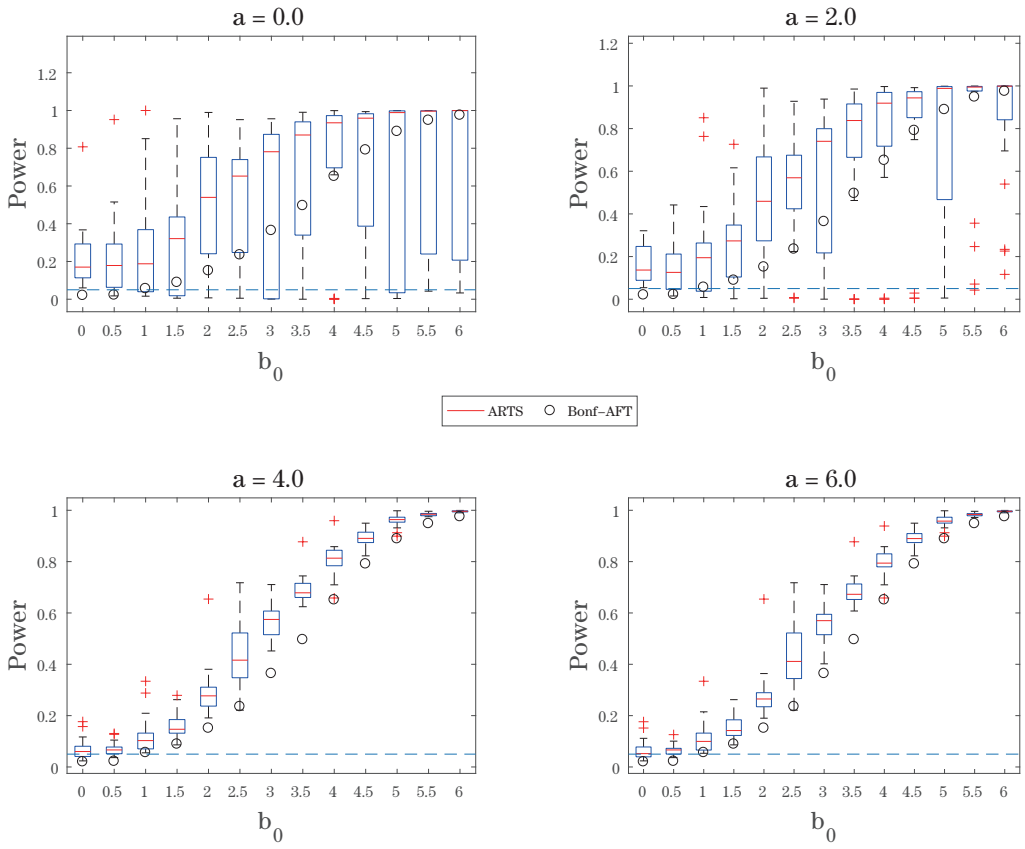

Figure 4. Asymptotic Type-I error and power, as in Figure 3, except under moderate censoring.

other. The FWERs of other AFT-model-relevant methods are also provided; here we omit the anti-conservative results of the CPB-AFT for conciseness, focusing instead on the CEND, which requires independence between ε and $\boldsymbol{U}$.

To produce a dependent error structure on the predictors, we generate the error term ε by random replications from a normal distribution with mean zero and a standard deviation of $0.7(|U_1| + 0.7)$. Then, we simulate the transformed time-to-event outcome under the null model $T = \varepsilon$. Though not independent, ε remains uncorrelated with $\boldsymbol{U}$ by $\mathrm{Cov}(\varepsilon, U_1) = E[\varepsilon U_1] = E\{U_1 E[\varepsilon|U_1]\} = 0$, and $\mathrm{Cov}(\varepsilon, U_j) = E\{U_j E[\varepsilon|U_1]\} = 0$ for $j \neq 1$. The censoring time C still follows an exponential distribution, with varying rate parameters specified for different censoring rates. Figure 6 shows that only the ARTS controls the FWER around the nominal level in the case of dependent errors, except for giving slightly

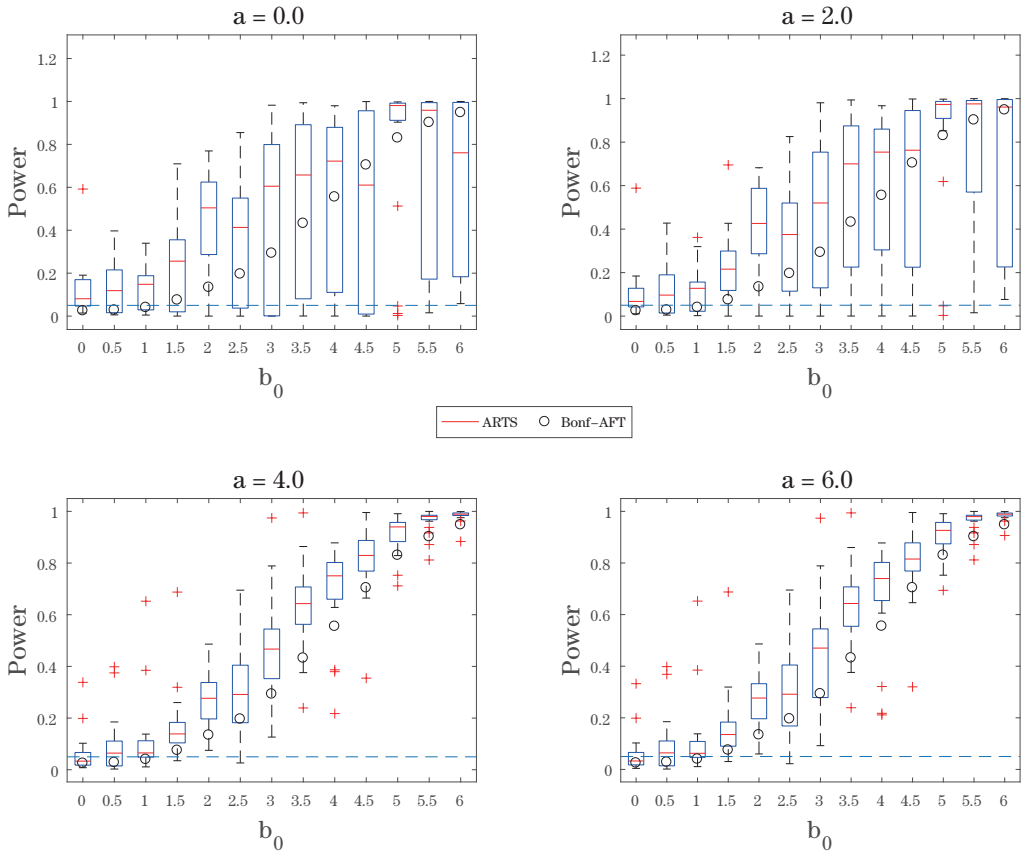

Figure 5. Asymptotic Type-I error and power, as in Figure 3, except under heavy censoring.

conservative FWERs for $p \geq 50$, heavy censoring, and $n = 100$.

7. Applications to Real Data

7.1. DLBCL data

We revisit the DLBCL data introduced earlier (Rosenwald et al. (2002)). This data set contains the after-chemotherapy survival time from DLBCL diseases, the categorical IPI variable (with three levels: low, medium, and high), and 7,399 genetic features of 222 patients with complete information on genetic predictors. The censoring rate is 43%. More details about the DLBCL data can be found in the literature (cf., Bøvelstad, Nygård and Borgan (2009), Binder, Porzelius and Schumacher (2011)). To adjust for the prognostic information pro-

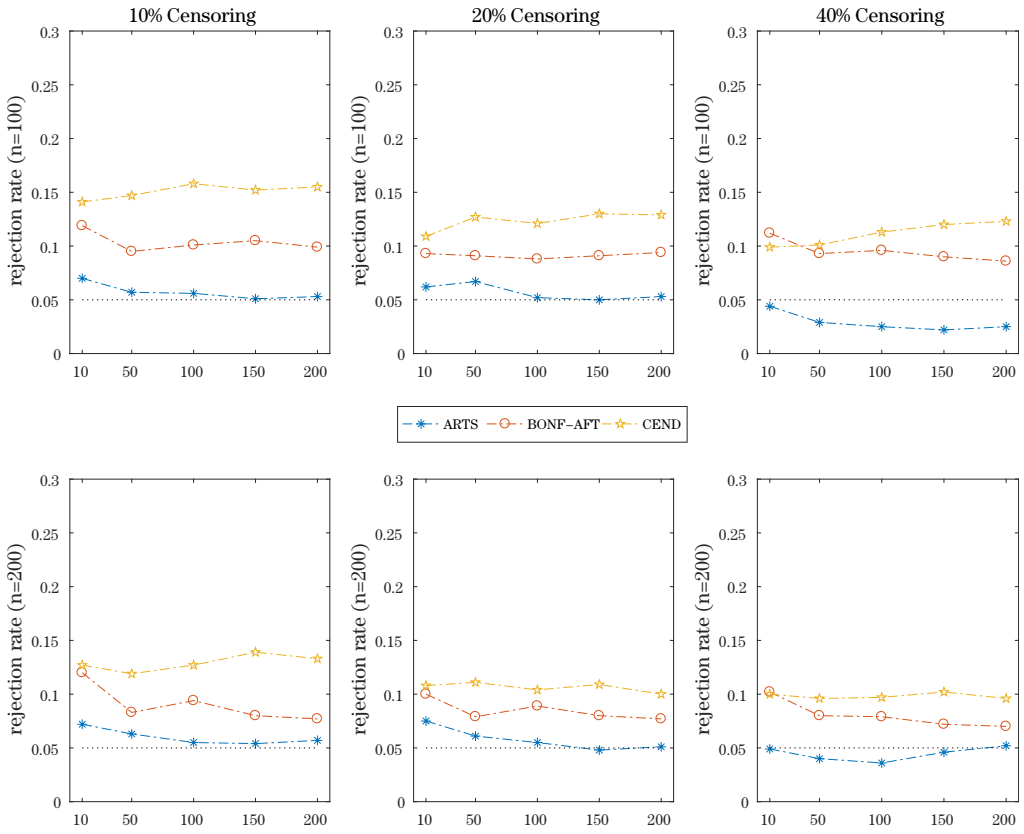

Figure 6. Empirical rejection rates based on 1,000 samples generated from the null model with dependent errors under various p, sample sizes, and censoring rates.

vided by IPI, we apply the adjusted ARTS to this data set to detect the presence of significant genetic features. To maintain the stability of the KSV estimator, the observed event times are restricted up to $\tau = 2.36$, which corresponds to the 90% empirical percentile of the observed event times. This excludes one observation that has an estimated synthetic response of 55.867 and severely distorts the estimation of the marginal regression coefficients. For the ARTS, we use the double-bootstrap to select the constant a from 0 to 15, by increments of 0.5. Before implementing the ARTS, we perform a pre-processing step to filter out genes that lack significant differentiation between the censored group (patients still alive at the end of the follow-up) and the uncensored group (patients who died of DLBCL diseases within the follow-up). For each gene, a standard two-sample t-test is conducted to determine whether the gene-expression measurement differentiates between these two groups. By comparing the corresponding p-values with

the nominal level of 5%, this pre-processing step reduces the number of screening genetic features to 1,026 ($p = 1{,}026$).

To give a fair comparison with the ARTS, we also apply the following AFT-model-relevant competing methods: BONF-AFT and CPB-AFT, with IPI information adjusted. The CEND method is not included, because it is challenging to verify the required assumption of independence between the error and the predictors. In addition, the HC method is not considered because it is designed for nearly uncorrelated predictors, which is unrealistic in gene-expression data. The three implemented approaches yield similar p-values. The minimal Bonferroni corrected p-value from the BONF-AFT is 4.39%. The ARTS procedure reduces to a special case with $\lambda_n = 0$ and gives the same p-value of 3.40% as that of the CPB-AFT, from 1,000 bootstrap samples. Figure 7 shows the sampling distribution of the test statistics used by the ARTS and CPB-AFT based on these bootstrap samples, as well as how the corresponding p-values are obtained. Given the nominal level of 5%, these three approaches all indicate one significant gene for the survival time of patients. The ID of the detected gene is "27,766," which belongs to the group of major histocompatibility class (MHC) II signatures. This finding supports the notion that a loss of MHC II expression correlates with a worse survival outcome, and corresponds to the results provided by Miller et al. (1988), Rosenwald et al. (2002), Rimsza et al. (2004), Roberts et al. (2006), and Higashi et al. (2016), among others.

7.2. Primary biliary cirrhosis data

In this example, we demonstrate how to apply the forward-stepwise ARTS to successively identify interaction effects, provided that the main effects of some covariates have been shown statistically or clinically significant. We use data from the Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984 (Fleming and Harrington (1991), Appendix D.1). A total of 312 PBC patients participated in the randomized placebo controlled trial of the drug D-penicillamine; in our data analysis, we restrict our attention to the 276 patients for whom we have complete covariate information. The censoring rate is 60%.

The survival outcome is the time from registration to death. Over the follow-up, there is no significant treatment effect (Fleming and Harrington (1991)). Only five of the 16 risk factors were found to be statistically significant under the setting of the Cox model (Dickson et al. (1989)) or under the AFT model (Jin et al. (2003)). Furthermore, they were identified as a subset of the active

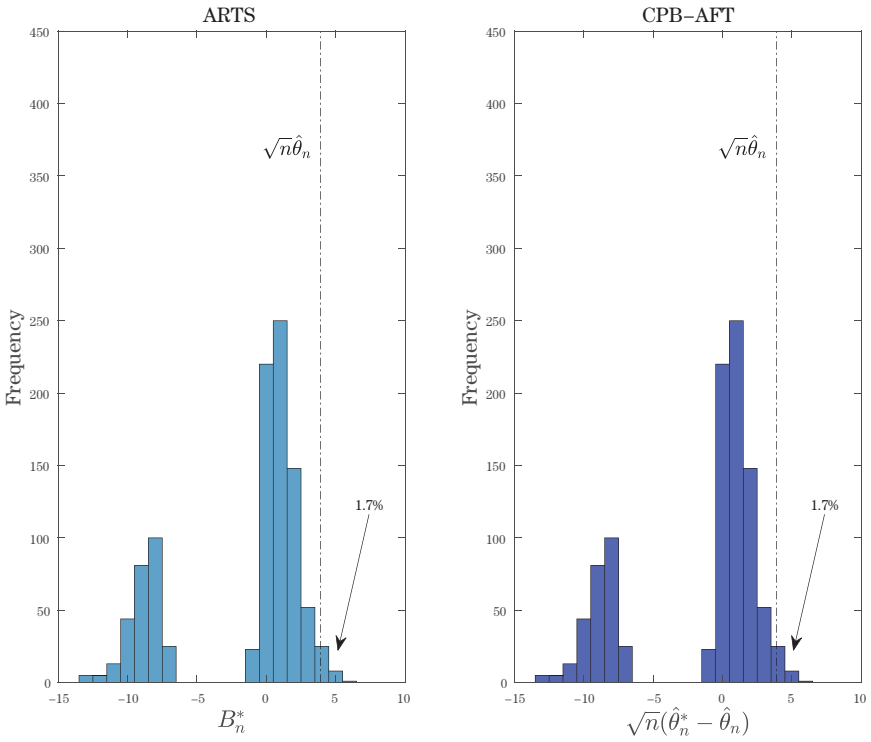

Figure 7. DLBCL example. Left panel: histogram of B_n^*, giving the two-sided ARTS p-value 3.40%. Right panel: histogram of $\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)$, giving the two-sided CPB-AFT p-value 3.40%.

predictors under the general Cox model (Bunea and McKeague (2005)). These significant risk factors are age (in years), presence of edema (0 = no; 0.5 = resolved; 1 = unresolved with therapy), serum bilirubin (in mg/dl), albumin (in gm/dl), and protime (standardized blood clotting time, in seconds). Of these risk factors, serum bilirubin, albumin, and protime are log-transformed. We successively locate significant pairwise interaction terms of 17 variables, adjusting for the five aforementioned risk factors. These 17 variables include the treatment indicator and 16 clinical risk factors for the survival time ($p = \binom{17}{2} = 136$).

Figure 8 displays the pattern of p-values for the newly entered interaction term at each step. The forward-stepwise ARTS procedure detects one significant interaction term, where the constant a and the end of the follow-up τ are selected as in Section 7.1. This detected interaction is between platelet (platelets per cubic ml/1,000) and alk.phos (alkaline phosphatase, in U/liter). For comparison, we also present the successive p-values given by the CPB-AFT. The conclusion

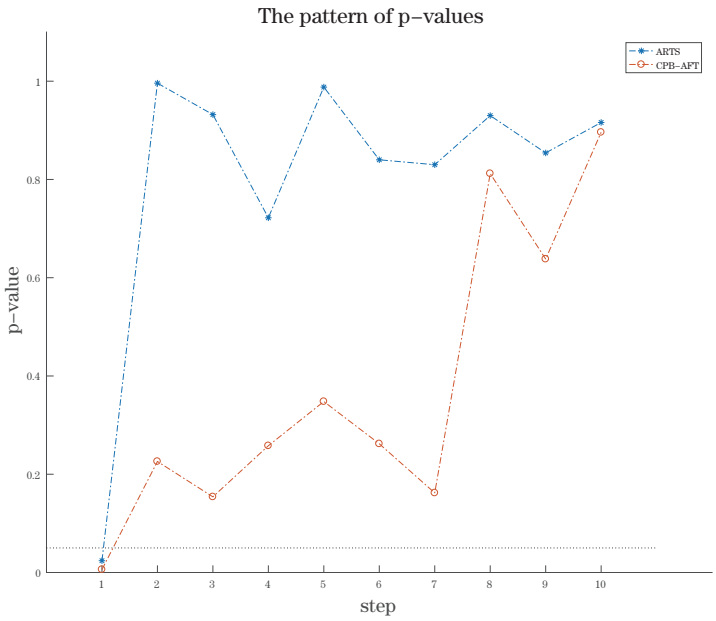

Figure 8. PBC example. The patterns of p-values for forward-stepwise ARTS and CPB-AFT.

remains the same, but the p-values of the CPB-AFT are smaller, as expected.

To examine the effect of taking covariate-dependent censoring into account when applying the ARTS in this example, we run the forward-stepwise ARTS as before, except we replace $\hat{G}_n$ by a Cox-model-based estimate, conditional on selected covariates (alkaline phosphotase and log-transformed protime). In contrast to our earlier finding of one significant interaction term, here we find none (results not shown). The CPB-AFT procedure (with the same Cox model estimate of G_0) leads to the same conclusion.

8. Discussion

We have developed an adaptive resampling test for survival data (ARTS) to detect the presence of significant predictors for right-censored survival outcomes. We use marginal correlation screening to reduce the high-dimensional detection problem to a single test of whether $\theta_0 = 0$, where θ_0 is the marginal regression coefficient of the most correlated predictor with the survival outcome. In the setting of marginal screening for survival data, few studies have examined the problem of post-selection inference. The problem is challenging, not only because of the nonregular asymptotic behavior of the test statistic at the

null (i.e., $\theta_0 = 0$), but also because of the presence of censoring. Within this framework, the ARTS is designed to adapt to the nonregularity, while dealing with the increased dispersion introduced by the censoring. The advantage of the ARTS is that it provides a post-selection-corrected p-value without sacrificing power, while avoiding distributional assumptions, specific correlation structures between predictors, and a preconceived choice of the regression parameters of interest. The ARTS procedure is also versatile for practical use. Various extensions of the ARTS are proposed to adjust for additional baseline covariates of clinicians' interests and to successively identify further active predictors.

We recognize that the ARTS requires an independent-censoring assumption that may be violated in some clinical contexts. One direction for future work is to develop rigorous theoretical results for the ARTS under the assumption of conditionally independent censoring, given the predictors. To address this type of censoring mechanism, we can use the Cox model or the local Kaplan–Meier estimator to incorporate covariates into the estimation of the conditional survival function of the censoring on the predictors $G_0(\cdot|\boldsymbol{U})$. The generalization of the censoring mechanism could still be challenging in our framework, even with some of the proposals for estimating $G_0(\cdot|\boldsymbol{U})$ listed above. One challenge is how to determine the covariates to be included in the estimation of $G_0(\cdot|\boldsymbol{U})$ under the high-dimensional AFT model. Then, we need to find out whether the post-selection inference results would be affected, because these included covariates may not be completely contained under a series of working AFT models using one predictor per time. To the best of our knowledge, this question has not been fully answered in the area of marginal screening based on survival data, and is worth further attention.

Although our simulation results show that the ARTS performs well when $p \gg n$, we have provided theoretical support only, assuming a fixed p. Formal testing procedures that can adjust to the nonregular behavior of $\hat{\theta}_n$ under diverging p appear to be challenging. A potential alternative approach that might be able to handle a diverging p would be to extend the efficient influence function technique of Luedtke and van der Laan (2018) to the right-censored setting in terms of a regularized version of the KSV estimator.

Supplementary Materials

The online Supplementary Material includes detailed proofs of the theorems, as well as additional simulation results.

Acknowledgment

This research was partially supported by NIH Grants R01GM095722 and R21MH108999, and NSF Grant DMS-1307838. The authors thank the associate editor and reviewers for their helpful comments.

References

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *The Annals of Statistics* **10**, 1100–1120.

Antoniadis, A., Fryzlewicz, P. and Letué, F. (2010). The Dantzig selector in Cox's proportional hazards model. *Scandinavian Journal of Statistics* **37**, 531–552.

Bender, R., Austin, T. and Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine* **24**, 1713–1723.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **57**, 289–300.

Binder, H., Porzelius, C. and Schumacher, M. (2011). An overview of techniques for linking high-dimensional molecular data to time-to-event endpoints by risk prediction models. *Biometrical Journal* **53**, 170–189.

Bøvelstad, H. M., Nygård, S. and Borgan, Ø. (2009). Survival prediction from clinico-genomic models–A comparative study. *BMC Bioinformatics* **10**, Article 413.

Bradic, J., Fan, J. and Jiang, J. (2011). Regularization for Cox's proportional hazards model with NP-dimensionality. *The Annals of Statistics* **39**, 3092–3120.

Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika* **66**, 429–436.

Bunea, F. and McKeague, I. W. (2005). Covariate selection for semiparametric hazard function regression models. *Journal of Multivariate Analysis* **92**, 186–204.

Cai, T., Huang, J. and Tian, L. (2009). Regularized estimation for the accelerated failure time model. *Biometrics* **65**, 394–404.

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

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269–276.

Dabrowska, D. (1989). Uniform consistency of the kernel conditional Kaplan–Meier estimate. *The Annals of Statistics* **17**, 1157–1167.

Datta, S., Le-Rademacher, J. and Datta, S. (2007). Predicting patient survival from microarray data by accelerated failure time modeling using partial least squares and LASSO. *Biometrics* **63**, 259–271.

Dickson, E. R., Grambsch, P. M., Fleming, T. R., Fisher, L. D. and Langworthy, A. (1989). Prognosis in primary biliary cirrhosis: model for decision making. *Hepatology* **10**, 1–7.

Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics* **32**, 962–994.

Donoho, D. and Jin, J. (2015). Higher criticism for large-scale inference, especially for rare and weak effects. *Statistical Science* **30**, 1–25.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap (Monographs on Statis-*

tics & Applied Probability). Chapman and Hall/CRC.

Engler, D. and Li, Y. (2009). Survival analysis with high-dimensional covariates: An application in microarray studies. *Statistical Applications in Genetics and Molecular Biology* **8**, Article 14.

Fan, J., Feng, Y. and Wu, Y. (2010). High-dimensional variable selection for Cox's proportional hazards model. *Borrowing Strength: Theory Powering Applications – A Festschrift for Lawrence D. Brown. Institute of Mathematical Statistics; Beachwood OH* **6**, 70–86.

Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics* **30**, 74–99.

Fang, E. X., Ning, Y. and Liu, H. (2017). Testing and confidence intervals for high dimensional proportional hazards models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79**, 1415–1437.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis*. John Wiley & Sons, Inc.

Goeman, J. J., Oosting, J., Cleton-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 ultra-high dimensional features. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75**, 217–245.

He, X., Wang, L. and Hong, H. G. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *The Annals of Statistics* **41**, 342–369.

Higashi, M., Tokuhira, M., Fujino, S., Yamashita, T., Abe, K., Arai, E., Kizaki, M. and Tamaru, J.-I. (2016). Loss of HLA-DR expression is related to tumor microenvironment and predicts adverse outcome in diffuse large B-cell lymphoma. *Leukemia & Lymphoma* **57**, 161–166.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**, 65–70.

Hong, H. G., Chen, X., Christiani, D. C. and Li, Y. (2018). Integrated powered density: screening ultra-high dimensional covariates with survival outcomes. *Biometrics* **74**, 421–429.

Hong, H. G., Kang, J. and Li, Y. (2018). Conditional screening for ultra-high dimensional covariates with survival outcomes. *Lifetime Data Analysis* **24**, 45–71.

Huang, J. and Ma, S. (2010). Variable selection in the accelerated failure time model via the bridge method. *Lifetime Data Analysis* **16**, 176–195.

Huang, J., Ma, S. and Xie, H. (2006). Regularized estimation in the accelerated failure time model with high-dimensional covariates. *Biometrics* **62**, 813–820.

Jin, Z., Lin, D. Y., Wei, L. J. and Ying, Z. (2003). Rank-based inference for the accelerated failure time model. *Biometrika* **90**, 341–353.

Johnson, B. A. (2008). Variable selection in semiparametric linear regression with censored data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**, 351–370.

Johnson, B. A., Lin, D. Y. and Zeng, D. (2008). Penalized estimating functions and variable selection in semiparametric regression models. *Journal of the American Statistical Association* **103**, 672–680.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Wiley Series in Probability and Statistics. Hoboken, NJ, USA: John Wiley & Sons, Inc.

Keiding, N., Andersen, P. K. and Klein, J. P. (1997). The role of frailty models and accelerated

failure time models in describing heterogeneity due to omitted covariates. *Statistics in Medicine* **16**, 215–224.

Koul, H., Susarla, V. and Van Ryzin, J. (1981). Regression analysis with randomly right-censored data. *The Annals of Statistics* **9**, 1276–1288.

Lai, T. L. and Ying, Z. (1991a). Large sample theory of a modified Buckley-James estimator for regression analysis with censored data. *The Annals of Statistics* **19**, 1370–1402.

Lai, T. L. and Ying, Z. (1991b). Rank regression methods for left-truncated and right-censored data. *The Annals of Statistics* **19**, 531–556.

Le Cessie, S. and van Houwelingen, H. C. (1995). Testing the fit of a regression model via score tests in random effects models. *Biometrics* **51**, 600–614.

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

Li, Y., Dicker, L. and Zhao, S. D. (2014). The dantzig selector for censored linear regression models. *Statistica Sinica* **24**, 251–268.

Lin, D. Y. and Ying, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika* **81**, 61–71.

Lockhart, R., Taylor, J., Tibshirani, R. J. and Tibshirani, R. (2014). A significance test for the lasso. *The Annals of Statistics* **42**, 413–468.

Luedtke, A. R. and van der Laan, M. J. (2018). Parametric-rate inference for one-sided differentiable parameters. *Journal of American Statistical Association* **113**, 780–788.

Ma, S. and Du, P. (2012). Variable selection in partly linear regression model with diverging dimensions for right censored data. *Statistica Sinica* **22**, 1003–1020.

McKeague, I. W. and Qian, M. (2015). An adaptive resampling test for detecting the presence of significant predictors (with discussion). *Journal of the American Statistical Association* **110**, 1422–1433.

McKeague, I. W. and Sasieni, P. D. (1994). A partly parametric additive risk model. *Biometrika* **81**, 501–514.

Medeiros, F. M., da Silva-Júnior, A. H., Valença, D. M. and Ferrari, S. L. (2014). Testing inference in accelerated failure time models. *International Journal of Statistics and Probability* **3**, 121–131.

Miller, T. P., Lippman, S. M., Spier, C. M., Slymen, D. J. and Grogan, T. M. (1988). HLA-DR (Ia) immune phenotype predicts outcome for patients with diffuse large cell lymphoma. *The Journal of Clinical investigation* **82**, 370–372.

Pollard, D. (1990). *Empirical Processes: Theory and Applications*. NSF-CBMS Regional Conference Series in Probability and Statistics: Vol 2. Institute of Mathematical Statistics.

Rimsza, L. M., Roberts, R. A., Miller, T. P., Unger, J. M., LeBlanc, M., Braziel, R. M., Weisenberger, D. D., Chan, W. C., Muller-Hermelink, H. K., Jaffe, E. S., Gascoyne, R. D., Campo, E., Fuchs, D. A., Spier, C. M., Fisher, R. I., Delabie, J., Rosenwald, A., Staudt, L. M. and Grogan, T. M. (2004). Loss of MHC class II gene and protein expression in diffuse large B-cell lymphoma is related to decreased tumor immunosurveillance and poor patient survival regardless of other prognostic factors: a follow-up study from the leukemia and lymphoma molecular profiling project. *Blood* **103**, 4251–4258.

Ritov, Y. (1990). Estimation in linear regression with censored data. *The Annals of Statistics* **18**, 303–328.

Roberts, R. A., Wright, G., Rosenwald, A. R., Jaramillo, M. A., Grogan, T. M., Miller, T. P.,

Frutiger, Y., Chan, W. C., Gascoyne, R. D., Ott, G., Muller-Hermelink, H. K., Staudt, L. M. and Rimsza, L. M. (2006). Loss of major histocompatibility class II gene and protein expression in primary mediastinal large B-cell lymphoma is highly coordinated and related to poor patient survival. *Blood* **108**, 311–318.

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, H. K., Smeland, E. B., Giltnane, J. M., Hurt, E. M., Zhao, H., Averett, L., Yang, L., Wilson, W. H., Jaffe, E. S., Simon, R., Klausner, R. D., Powell, J., Duffey, P. L., Longo, D. L., Greiner, T. C., Weisenburger, D. D., Sanger, W. G., Dave, B. J., Lynch, J. C., Vose, J., Armitage, J. O., Montserrat, E., Lpez-Guillermo, A., Grogan, T. M., Miller, T. P., LeBlanc, M., Ott, G., Kvaloy, S., Delabie, J., Holte, H., Krajci, P., Stokke, T. and Staudt, L. M. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *The New England journal of medicine* **346**, 1937–1947.

Sinnott, J. A. and Cai, T. (2016). Inference for survival prediction under the regularized Cox model. *Biostatistics* **17**, 692–707.

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

Srinivasan, C. and Zhou, M. (1994). Linear regression with censoring. *Journal of Multivariate Analysis* **49**, 179–201.

Stute, W. and Wang, J.-L. (1993). The strong law under random censorship. *The Annals of Statistics* **21**, 1591–1607.

Taylor, J. and Tibshirani, R. (2018). Post-selection inference for ℓ_1-penalized likelihood models. *Canadian Journal of Statistics* **46**, 41–61.

The International Non-Hodgkin's Lymphoma Prognostic Factors Project (1993). A predictive model for aggressive non-Hodgkin's lymphoma. *New England Journal of Medicine* **329**, 987–994.

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

Tsiatis, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. *The Annals of Statistics* **18**, 354–372.

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.

Wu, Y. (2012). Elastic net for Cox's proportional hazards model with a solution path algorithm. *Statistica Sinica* **22**, 271–294.

Ying, Z. (1993). A large sample study of rank estimation for censored regression data. *The Annals of Statistics* **21**, 76–99.

Zhang, H. H. and Lu, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika* **94**, 691–703.

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.

Zhao, S. D. and Li, Y. (2014). Score test variable screening. *Biometrics* **70**, 862–871.

Zhong, P.-S., Hu, T. and Li, J. (2015). Tests for coefficients in high-dimensional additive hazard models. *Scandinavian Journal of Statistics* **42**, 649–664.

Zhou, M. (1992). Asymptotic normality of the 'synthetic data' regression estimator for censored survival data. *The Annals of Statistics* **20**, 1002–1021.

Department of Biostatistics, Columbia University, New York, NY 10027, USA.
E-mail: th2455@caa.columbia.edu
Department of Biostatistics, Columbia University, New York, NY 10027, USA.
E-mail: im2131@cumc.columbia.edu
Department of Biostatistics, Columbia University, New York, NY 10027, USA.
E-mail: mq2158@cumc.columbia.edu

(Received July 2017; accepted February 2018)