Information criteria for detecting change-points in the Cox proportional hazards model

Ryoto Ozaki\textsuperscript{1,2} | Yoshiyuki Ninomiya\textsuperscript{2,3}

\textsuperscript{1}Biometrics Department, Chugai Pharmaceutical Co., Ltd., Tokyo, Japan
\textsuperscript{2}Department of Statistical Science, The Graduate University for Advanced Studies, Tokyo, Japan
\textsuperscript{3}Department of Statistical Inference and Mathematics, The Institute of Statistical Mathematics, Tokyo, Japan

Correspondence
Yoshiyuki Ninomiya, Department of Statistical Inference and Mathematics, The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa-shi, Tokyo 190-8562, Japan.
Email: ninomiya@ism.ac.jp

Funding information
JSPS Grant-in-Aid for Scientific Research, Grant/Award Number: 16K00050

Abstract
The Cox proportional hazards model, commonly used in clinical trials, assumes proportional hazards. However, it does not hold when, for example, there is a delayed onset of the treatment effect. In such a situation, an acute change in the hazard ratio function is expected to exist. This paper considers the Cox model with change-points and derives Akaike information criterion (AIC)-type information criteria for detecting those change-points. The change-point model does not allow for conventional statistical asymptotics due to its irregularity, thus a formal AIC that penalizes twice the number of parameters would not be analytically derived, and using it would clearly give overfitting analysis results. Therefore, we will construct specific asymptotics using the partial likelihood estimation method in the Cox model with change-points, and propose information criteria based on the original derivation method for AIC. If the partial likelihood is used in the estimation, information criteria with penalties much larger than twice the number of parameters could be obtained in an explicit form. Numerical experiments confirm that the proposed criteria are clearly superior in terms of the original purpose of AIC, which are to provide an estimate that is close to the true structure. We also apply the proposed criterion to actual clinical trial data to indicate that it will easily lead to different results from the formal AIC.

KEYWORDS
Brownian motion, model misspecification, model selection, statistical asymptotic theory, structural change, survival time analysis

1 INTRODUCTION

The proportional hazards model proposed by Cox (1972) is widely used for survival analysis in clinical studies with time-to-event endpoints. This model involves potential assumptions such as proportional hazards and log-linearity. Proportional hazards assume that the hazard ratio is constant over time, but the data structure in actual clinical trials often deviates from this assumption. Various problems have been pointed out here: the lack of theoretical validity in applying analytical methods for which the proportional hazards property is assumed in situations where it does not hold; the reduced detection power for covariates; and the difficulty of interpreting estimated hazard ratios (see, e.g., Uno et al., 2014). The problem of deviations from the proportional hazards assumption has been around for a long time, and works such as Gill and Schumacher (1987) and Hess (1995)
proposed methods to detect and avoid deviations from the assumption.

In clinical trials comparing therapeutic products such as immune checkpoint inhibitors or cancer vaccines with control products such as placebos, it takes time from the start of treatment to the onset of efficacy because of the products’ mode of action. The survival curves in such clinical trials are expected to overlap for a certain period after the start of treatment, and then the difference between the curves increases. As well as these delayed effect cases, there are also cases in which the proportional hazard property does not hold, such as cases in which the curves that have diverged run parallel after a while, or cases in which the curves converged rather than run parallel. Because these are considered to be one or more acute changes in the hazard ratio function, change-point analysis is important for detecting such time points in survival time analysis. Liu et al. (2008) and He et al. (2013) proposed methods based on the maximum score test and a sequential test approach, respectively, to detect change-points in a hazard ratio that is assumed to be constant in segmental terms. However, there are still no reported methods for detecting change-points by using information criteria. Moreover, especially in nonrandomized observational studies or in cases where the number of covariates is large relative to the sample size, there is often multicollinearity among covariates, which induces unstable estimation. Therefore, ridge regression has been used as an alternative estimation method in survival time analyses based on the Cox model (see, e.g., van Houwelingen et al., 2006 and Witten & Tibshirani, 2010).

Accordingly, in this paper, we use a combination of the Cox model and a change-point model to develop information criteria that are reasonable in the sense that they are based on the original derivation of the Akaike information criterion (AIC, Akaike, 1973), that is, they are obtained as asymptotically unbiased estimators of properly defined risks. Specifically, we use the partial likelihood and add a regularization term to the loss function used for estimation. Then, from the viewpoint of the conventional AIC, we assume the existence of \( m \) change-points in developing the AIC for a model with \( m \) change-points. Our derivation of asymptotically unbiased estimators is theoretically based on Tsaiitis (1981), which derived the asymptotic properties of regression parameter estimators via the partial likelihood method, and Pons (2002), which evaluated asymptotic theories in a change-point model.

Because a change-point model requires specific asymptotic theories (see, e.g., Csörgő & Horváth, 1997), various theories have been developed for the test-based approach, which implies that the information criteria for a model with change-points also require specific theories. Siegmund (2004) first derived an information criterion based on the original definition by using theory specific to the change-point model; that is, the approach was reasonable in terms of original derivation rather than formal. Hence, the purpose of this paper is to derive AIC-type information criteria. For a basic change-point model in which data are observed independently at each time, such a derivation was performed by Ninomiya (2015). That work gave the criterion as an asymptotically unbiased estimator of the Kullback–Leibler divergence between the true and estimated distributions. According to that derivation, the asymptotic bias due to the regression parameter is 1, whereas that due to the change-point parameter is 3, that is, three times larger. This implies that consideration of theory specific to a change-point model in the AIC derivation would significantly alter the analytical results.

The Cox model with change-points can be regarded as a model that assumes that proportional hazards do not hold for the entire follow-up period but do hold for segmented periods. We also assume log-linearity in each interval, as in the conventional Cox model. This log-linearity implies that the hazard function yields a linear expression when the logarithm of its covariate-dependent part is taken. In observational studies, it is rare to measure all the covariates that are necessary for hazard ratio estimation; moreover, even if log-linearity is established, the model used for estimation may be misspecified. Of course, there is a deviation from log-linearity when the effect of covariates cannot be expressed by a linear equation. In addition, if the regression parameters are estimated by the conventional partial likelihood method in such a case, the estimator will not have good properties such as consistency, which makes it difficult to interpret the hazard ratios. Estimation under conditions of model misspecification in the Cox model was discussed in Struthers and Kalbfleisch (1986) and Lin and Wei (1989). In this paper, by leveraging those discussions, we also extend the proposed AIC to handle such conditions of model misspecification, in the form of the Takeuchi information criterion (TIC, Takeuchi, 1976).

We note here that in the Cox model with change-points, unlike the usual change-point model, the time and outcome variables are not observed separately, but both are the same variable of survival time. When the time is not observed, or when both are the same but with added noise, this change-point model will have the same type of structure as the so-called mixture distribution models. As described in Dacunha-Castelle and Gassiat (1999), mixture distribution models also have asymptotic properties that are different from those of regular statistical models and even more different from those of change-point models. In other words, it is not at all obvious whether the AIC addressed in this paper is of the same type as the AIC for a conventional change-point model or the AIC for a mixture distribution model. The information criteria proposed in this paper should thus clarify this point.
We organized the rest of this paper as follows. In Section 2, for preliminaries, we first define the Cox model with change-points and assume the usual conditions for asymptotic theories in change-point analysis. Then, by using the original AIC derivation, we define an AIC-type information criterion as an asymptotically bias-corrected version of the regularized maximum log-partial likelihood. In Section 3, we evaluate the asymptotic bias and show that it can be expressed explicitly. To evaluate our approach, we describe the results of numerical experiments in Section 4. First, we confirm that the asymptotic bias evaluation does indeed approximate the bias accurately. Then, the performance of the derived AIC is compared with that of the formally defined AIC without using any theory specific to the change-point model. For further evaluation, in Section 5, we describe a case study on applying the derived AIC and the formally defined AIC to actual clinical trial data for change-point detection. In Section 6, we extend our theory and proposed AIC to the case where model misspecification is allowed, and we derive TIC. Finally, we give our conclusion in Section 7.

2 | PRELIMINARIES

2.1 | Model and assumptions

For the Cox model with shifts in the regression parameters, we incorporate a model with \( m \) change-points as follows:

\[
\lambda(t \mid z) = \lambda_0(t) \exp(\beta^{(j)^T} z), \quad t \in [k^{(j-1)}, k^{(j)}),
\]

\( j \in \{1, 2, \ldots, m+1\}. \tag{1} \]

Here, \( \lambda(t \mid z) \) is the hazard function under a given covariate vector \( z \), and that \( \lambda_0(t) \) is the baseline hazard function. For each \( j \in \{1, 2, \ldots, m+1\} \), let the regression parameter \( \beta^{(j)} \) be a \( p \)-dimensional vector; that is, \( \beta \equiv (\beta^{(1)^T}, \beta^{(2)^T}, \ldots, \beta^{(m+1)^T})^T \) is a \( (m+1) \)-dimensional vector. Let \( k \equiv (k^{(1)}, k^{(2)}, \ldots, k^{(m)})^T \) be an \( m \)-dimensional vector of change-point parameters, and let \( k^{(0)} = 0 \) and \( k^{(m+1)} = T \), where \( T \) is the follow-up period of the survival time. Also, suppose that the true values of \( k \) and \( \beta \) are \( k^* = (k^{*(1)}, k^{*(2)}, \ldots, k^{*(m)})^T \) and \( \beta^* = (\beta^{*(1)^T}, \beta^{*(2)^T}, \ldots, \beta^{*(m+1)^T})^T \), respectively, such that \( 0 < k^{*(1)} < k^{*(2)} < \cdots < k^{*(m)} < T \). To construct a change-point model, we assume that

\[
\beta^{*(1)} \neq \beta^{*(2)} \neq \cdots \neq \beta^{*(m+1)}, \tag{2}
\]

and that \( k^* \) and \( \beta^* \) are unknown. For simplicity, let the space of \( \beta \) be compact.

The regression parameters in the Cox model are usually estimated by the partial likelihood method proposed by Cox (1972). First, let \( y_1 \) and \( y_2 \) be positive random variables that denote the times of an event and a censoring occurrence, respectively. We assume that \( y_1 \) and \( y_2 \) are conditionally independent given the covariate vector \( z \). The time to finish an observation, that is, the time to an event or censoring, can be expressed as \( t = \min(y_1, y_2) \). From \( y_1 \) and \( y_2 \), we define \( \delta \) as a random variable taking a value of 1 for \( y_1 \leq y_2 \) (event) and 0 for \( y_1 > y_2 \) (censoring). For an experiment with \( n \) subjects, let \( t \equiv (t_1, t_2, \ldots, t_n) \) and \( D = (k^{(j-1)}_1, k^{(j)}_1) \equiv \{ i \mid \delta_i = 1, t_i \in [k^{(j-1)}, k^{(j)}): \ j = 1, 2, \ldots, n \} \) denote the vector of time to event or censoring and the set of subjects for which the event occurs in the period \( [k^{(j-1)}, k^{(j)}) \), respectively. Then, the partial likelihood function is given by

\[
L(\beta, k; t) \equiv \prod_{j=1}^{m+1} \prod_{i \in D([k^{(j-1)}, k^{(j)}))} \frac{\exp(\beta^{(j)^T} z_i)}{\sum_{t' \in E(t_i)} \exp(\beta^{(j)^T} z_{t'})},
\]

where \( R(t) \) is the risk set \( \{ i \mid t < t_i \} \) at time \( t \). Also, we can express the log-partial likelihood function as

\[
l(\beta, k; t) \equiv \sum_{j=1}^{m+1} \sum_{i \in D([k^{(j-1)}, k^{(j)}))} \left[ \beta^{(j)^T} z_i - \log \left( \sum_{t' \in E(t_i)} \exp(\beta^{(j)^T} z_{t'}) \right) \right]. \tag{3}
\]

The approach in Cox (1972) estimates the parameters \( \beta \) and \( k \) by maximizing this log-partial likelihood function. However, we generalize that approach here to estimate the regression and change-point parameters by maximizing the regularized log-partial likelihood function with a ridge-type regularization term (Hoerl & Kennard, 1970) for the model given by (1). Specifically, we define the regularized log-partial likelihood function as

\[
l_1(\beta, k; t) \equiv \sum_{j=1}^{m+1} \sum_{i \in D([k^{(j-1)}, k^{(j)}))} \left[ \beta^{(j)^T} z_i - \log \left( \sum_{t' \in E(t_i)} \exp(\beta^{(j)^T} z_{t'}) \right) \right] - \frac{\xi}{2} \beta^{(j)^T} \beta^{(j)}, (3)\]

and we estimate the parameters by maximizing this function, where \( \xi \) is a regularization parameter. Hereafter, we denote \( D(\gamma) \equiv \{ k^{(j-1)}, k^{(j)} \} \) to simplify the notation.

Let \( \hat{\beta}_k \) be the \( \beta \) that maximizes the regularized log-partial likelihood function with the fixed change-points \( k \), that is, \( \hat{\beta}_k \equiv \arg\sup_{\beta} l_1(\beta, k; t) \). Also, we denote \( \hat{k} \equiv \arg\sup_k l_1(\hat{\beta}_k, k; t) \) as the estimator of the change-point parameter. Then, the regression parameter estimator is given by \( \hat{\beta} \equiv \hat{\beta}_k \).

Next, we define \( h(t_1, \beta^{(j)}_1) \) and \( H(t_1, \beta^{(j)}_1) \) as a \( p \times p \) matrix

\[
\left. \sum_{t' \in E(t_i)} \exp(\beta^{(j)^T} z_{t'}) \right/ \sum_{t' \in E(t_i)} \exp(\beta^{(j)^T} z_{t'})
\]

and a \( p \times p \) matrix \( \sum_{t' \in E(t_i)} z_i z_{t'}^T \).
\[
\exp(\beta_j^T z_{ij}) / \sum_{j' \in R(t_i)} \exp(\beta_{j'}^T z_{ij'}),
\]
respectively. Then, for each \(j \in \{1, 2, \ldots, m + 1\}\), the first and second derivatives on \(\beta_j^{(j)}\) for the regularized log-partial likelihood function can be expressed as follows:

\[
\frac{\partial}{\partial \beta_j^{(j)}} l_\xi(\beta, k; t) = \sum_{i \in B(j)} \{ z_i - h(t_i, \beta^{(j)}) - \xi \beta_j^{(j)} \},
\]
and

\[
\frac{\partial^2}{\partial \beta_j^{(j)} \partial \beta_j^{(j)}} l_\xi(\beta, k; t) = - \sum_{i \in B(j)} H(t_i, \beta^{(j)})
- h(t_i, \beta^{(j)}) h(t_i, \beta^{(j)})^T + \xi I_p,
\]
where \(I_p\) is the \(p\)-dimensional identity matrix.

### 2.2 Asymptotics for partial likelihood method

For the case when there are no change-points, Tsiatis (1981) showed the consistency and asymptotic normality of an estimator based on the partial likelihood method for the regression parameter \(\beta\) in the Cox model, where the score function for the partial likelihood is expressed as a sum of independent random variables. In this subsection, we reveal the asymptotic behavior of the estimator \(\hat{\beta}_k\) by maximizing the regularized log-partial likelihood based on (3). Because the possible range of times to events or censoring occurrences \(t_1, t_2, \ldots, t_n\) is \([0, T]\), the values of the change-points \(k_1, k_2, \ldots, k_m\) are finite even when considering the asymptotic theories as \(n\) increases. First, we define a \(p\)-dimensional square matrix as

\[
A_{\xi}^{(j)}(\beta, k) \equiv E\left\{ \frac{1}{n} \sum_{i \in B(j)} \{ z_i - h(t_i, \beta^{(j)}) - \xi \beta_j^{(j)} \} \right\}^T \left\{ \sum_{i \in B(j)} \{ z_i - h(t_i, \beta^{(j)}) - \xi \beta_j^{(j)} \} \right\},
\]
and denote \(A_{\xi}^{(j)} \equiv A_{\xi}^{(j)}(\beta^*, k^*)\) for simplicity, where \(\beta_{\xi}^* \equiv \text{arg sup}_{\beta} E[l_\xi(\beta, k^*; t)]\). Note that

\[
E[(\partial / \partial \beta_j^{(j)}) l_\xi(\beta_{\xi}^*, k^*; t)] = E[(\partial / \partial \beta_j^{(j)}) l_\xi(\beta_{\xi}^*, k^*; t)]
- \xi D_{\xi}^{(j)}(\beta_{\xi}^*) = 0_p, \quad \text{where } D_{\xi}^{(j)} = E[|D(\{k^{(j-1)}, k^{(j)})|/n]/n\text{ and } 0_p \text{ is a } p \text{-dimensional zero vector.}
\]

For a vector of finite values, \(s \equiv (s^{(1)}, s^{(2)}, \ldots, s^{(m)})^T\), we set \(k^{(j)} = k^{(j-1)} + s^{(j)}/n\) for each \(j \in \{1, 2, \ldots, m\}\). Then, similarly to Tsiatis (1981), we can trivially show the consistency and asymptotic normality of \(\hat{\beta}_k\), and we obtain

\[
\| \hat{\beta}_k - \beta_{\xi}^{(j)} \| = o_p(1)
\]
and

\[
\sqrt{n} \left( \hat{\beta}_k - \beta_{\xi}^{(j)} \right) \xrightarrow{d} N \left( 0_p, \left( A_{\xi}^{(j)} + \xi D_{\xi}^{(j)} I_p \right)^{-1} \right),
\]
where \(N(\mu, \Sigma)\) is a multivariate normal distribution with mean vector \(\mu\) and variance–covariance matrix \(\Sigma\).

### 2.3 AIC for partial likelihood method

In this subsection, we introduce the AIC for the Cox model when the partial likelihood is used, which was derived in Xu et al. (2009). They used a risk function based on the Kullback–Leibler divergence between the true and estimated models, as in the case of conventional AIC-type information criteria. The asymptotic bias for each of the estimated parameters has been shown to be 1. Later, we will derive AIC-type information criteria for the model given by (1) in the same way.

Let \((\hat{\beta}_i, \hat{k}_i) \equiv \text{arg sup}_{\beta, k} l_\xi(\beta, k; t)\) be an estimator of \((\beta, k)\) based on survival time data \(t \equiv (t_1, t_2, \ldots, t_n)^T\). In addition, by letting \(u \equiv (u_1, u_2, \ldots, u_n)^T\) be a copy of \(t\), that is, letting \(u\) independently follow the same distribution as \(t\), we obtain a divergence \(-2E_{\xi}[l_\xi(\hat{\beta}_i, \hat{k}_i; u)]\) based on the risk used in estimation, where \(E_u\) denotes the expectation with respect to \(u\). Then, for an initial estimator, we take \(-2\) times the maximum regularized log-partial likelihood \(-2l_\xi(\hat{\beta}_i, \hat{k}_i; t)\), which can be bias-corrected by

\[
E_u[2l_\xi(\hat{\beta}_i, \hat{k}_i; t) - E_u[2l_\xi(\hat{\beta}_i, \hat{k}_i; u)]
= 2E[l_\xi(\hat{\beta}_i, \hat{k}_i; t) - l_\xi(\hat{\beta}_u, \hat{k}_u; t)].
\]

However, as this expectation cannot be given explicitly, we will evaluate the bias asymptotically, as with the conventional AIC. First, by defining \(l_{\xi}(\beta, k; t) \equiv l_\xi(\beta, k; t) - l_\xi(\beta^*, k^*; t)\), \(\hat{\beta}_{k, u} \equiv \text{arg sup}_{\beta} l_{\xi}(\beta, k; u)\), and \(l_{\xi}(k; t, u) \equiv l_{\xi}(\hat{\beta}_{k, u}, k; t)\), we express the bias as

\[
2E[l_{\xi}(\hat{\beta}_t, \hat{k}_t; t) - l_{\xi}(\hat{\beta}_u, \hat{k}_u; t)]
= 2E \left[ \sup_k ll_{\xi}(\hat{\beta}_{k, t}, k; t) \right]
\]
\[-l_{\xi} \left\{ \hat{\theta}_{\text{argmax}_k} \ell_{\xi}(\hat{\theta}_{k,u}; \mathbf{k}; u) \right\} \leq 2E \left[ \sup_{k} \hat{\ell}_{\xi}(\mathbf{k}; t, t) - \hat{\ell}_{\xi}(\mathbf{k}; \mathbf{u}, u; t, u) \right], \]

Also, by defining \( b_{\xi}(\mathbf{k}^*, \hat{\beta}_{\xi}^*) \) as the weak limit of \( \sup_{k \in K} \hat{l}_{\xi}(\mathbf{k}; t, t) - \hat{l}_{\xi}(\text{argmax}_k \hat{l}_{\xi}(\mathbf{k}; u, \mathbf{u}; t, u)) \), we regard \( 2E[b_{\xi}(\mathbf{k}^*, \hat{\beta}_{\xi}^*)] \) as the asymptotic bias. Here, \( K \) denotes the set such that \( \hat{l}_{\xi}(\mathbf{k}; t, t) \) is \( \mathcal{O}(1) \) or positive; that is, it denotes the set for which there exists some positive constant \( M \) such that \( P[\hat{l}_{\xi}(\mathbf{k}; t, t) > -M] \) does not converge to 0. Then, we can say that

\[-2l_{\xi}(\hat{\beta}; t, i; t) + 2E[b_{\xi}(\mathbf{k}^*, \hat{\beta}_{\xi}^*)] \]

is the AIC for the Cox model with change-points when using the regularized partial likelihood method. If there are no change-points and the regularization parameter \( \xi \) is 0, then this is the same as the AIC given by Xu et al. (2009), where \( E[b_{\xi}(\mathbf{k}^*, \hat{\beta}_{\xi}^*)] = E[b_{\xi}(\mathbf{k}, \hat{\beta}_{\xi})] \) with \( \xi = 0 \) becomes the number of parameters in \( \hat{\beta} \).

## 3 MAIN RESULTS

In this section, under the setting of Section 2.1, we use the asymptotic property obtained in Section 2.2 to develop a novel information criterion by reevaluating the asymptotic bias according to the original AIC derivation method, which was introduced in Section 2.3.

### 3.1 Evaluation of asymptotic bias

Let us set \( k^{(j)} = k^{\text{st}} + s^{(j)}/n \) for each \( j \in \{1, 2, ..., m\} \). First, we consider the case where \( s = (s^{(1)}, s^{(2)}, ..., s^{(m)}) \) is a vector with finite values. Then, it follows from taking the difference between evaluations of \( \hat{\beta}_{\xi}^{(j)} - \hat{\beta}_{\xi}^{(j)} \) and \( \beta_{\xi}^{(j)} - \beta_{\xi}^{(j)} \) that

\[ \hat{\beta}_{\xi}^{(j)} - \hat{\beta}_{\xi}^{(j)} = \left[ \sum_{i \in D^{(j)}} (H(t_i, \beta_{\xi}^{(j)}) - h(t_i, \beta_{\xi}^{(j)})h(t_i, \beta_{\xi}^{(j)} + \xi I_p) \right]^{-1} \]

\[ \left\{ z_i - h(t_i, \beta_{\xi}^{(j)}) - \xi \beta_{\xi}^{(j)} \right\} \]

\[ - \sum_{i \in D^{(j)} \cap D^{(j)}} \left\{ z_i - h(t_i, \beta_{\xi}^{(j)}) - \xi \beta_{\xi}^{(j)} \right\} \]

\[ \{1 + o_p(1), \} \]

and therefore \( \| \hat{\beta}_{\xi}^{(j)} - \hat{\beta}_{\xi}^{(j)} \| = \mathcal{O}_{p}(1/n) \), where \( D^{(j)} \equiv D([k^{(j)-1}, k^{(j)})) \). See Web Appendix A for the derivation of (6). From Tsiatis (1981), we also have \( \| \hat{\beta}_{\xi}^{(j)} - \beta_{\xi}^{(j)} \| = \mathcal{O}_{p}(1/\sqrt{n}) \), which implies that (4) holds. Next, by using Taylor expansion around \( \beta_{\xi}^{(j)} = \hat{\beta}_{\xi}^{(j)} \), for the regularized log-partial likelihood function (6), we have

\[ l_{\xi}(\hat{\beta}; t, i; t) - l_{\xi}(\beta_{\xi}^{(j)}, k^*; t) \]

\[ = \sum_{j=1}^{m+1} \left( -\beta_{\xi}^{(j)} - \hat{\beta}_{\xi}^{(j)} \right) \left\{ \sum_{i \in D^{(j)}} \{ z_i - h(t_i, \beta_{\xi}^{(j)}) - \xi \beta_{\xi}^{(j)} \} \right\} \]

\[ + \frac{1}{2} \left( \beta_{\xi}^{(j)} - \hat{\beta}_{\xi}^{(j)} \right) \left[ \sum_{i \in D^{(j)}} \{ H(t_i, \beta_{\xi}^{(j)}) - h(t_i, \beta_{\xi}^{(j)})h(t_i, \beta_{\xi}^{(j)})^{T} \}ight] \]

\[ + \xi I_p \right\} (\beta_{\xi}^{(j)} - \hat{\beta}_{\xi}^{(j)}) \]

\[ + o_p(1) \]

\[ = \sum_{j=1}^{m+1} \left( \frac{1}{2} \left( \beta_{\xi}^{(j)} - \hat{\beta}_{\xi}^{(j)} \right) \right) \left[ \sum_{i \in D^{(j)}} \{ H(t_i, \beta_{\xi}^{(j)}) - h(t_i, \beta_{\xi}^{(j)})h(t_i, \beta_{\xi}^{(j)})^{T} \} \right] \]

\[ \left( \beta_{\xi}^{(j)} - \hat{\beta}_{\xi}^{(j)} \right) \]

\[ + o_p(1) \]

\[ = l_{\xi}(\hat{\beta}; t, i; t) - l_{\xi}(\beta_{\xi}^{(j)}, k^*; t) + o_p(1). \]
as a two-sided random walk with negative drift, we obtain

\[
l_2(\hat{\beta}_k, k; t) - l_2(\hat{\beta}_k^*, k^*; t) = l_2(\beta_0^*, k^*; t) - l_2(\hat{\beta}_k^*, k^*; t) + o_p(1) \\
= \sum_{j=1}^{m} Q^{(j)}_{\xi k^*+t/n,t} + o_p(1) = O_p(1). \tag{8}
\]

Furthermore, by using Taylor expansion around \( \hat{\beta}_k^* = \beta_0^* \) for the regularized log-partial likelihood, and from (4) and Murphy and van der Vaart (2000), the following holds:

\[
l_2(\hat{\beta}_k^*, k^*; t) - l_2(\beta_0^*, k^*; t) = \frac{1}{2} \sum_{j=1}^{m+1} \mathbf{v}^{(j)}_\xi \mathbf{v}^{(j)}_\xi + o_p(1) = O_p(1), \tag{9}
\]

where \( \mathbf{v}^{(j)}_\xi \) is a random variable vector distributed according to a multivariate normal distribution \( N(0, \mathbf{A}^{(j)}_\xi^{-1}(\mathbf{A}^{(j)}_\xi + \xi D^{(j)} I_p)) \). From (8) and (9), we thus have

\[
\hat{l}_2(\mathbf{k}; t, t) = O_p(1).
\]

Second, we consider the case where \( s \) is not a vector with finite values. From the discussion in Web Appendix B, in this case, we have \( P[\hat{l}_2(\hat{\beta}_k, k; t) - \hat{l}_2(\hat{\beta}_k^*, k^*; t) > -M] \to 0 \) for any \( M > 0 \); it then follows from (9) that \( P[\hat{l}_2(\mathbf{k}; t, t) > -M] \to 0 \) for any \( M > 0 \). Thus, we obtain

\[
\| \hat{k} - k^* \| = O_p(1/n), \tag{10}
\]

which is consistent with the result in Pons (2002).

From the above derivation, it can be seen that \( K = \{ k | k^{(j)} \in K^{(j)}, j \in \{1, 2, ..., m\} \} \), where \( K^{(j)} = \{ k | k - k^* = O(1/n) \} \). Therefore, from (8) and (9), we have

\[
\sup_{k \in K} \hat{l}_2(\mathbf{k}; t, t) = \sum_{j=1}^{m} \sup_{k \in K^{(j)}} Q^{(j)}_{\xi k,t} + \frac{1}{2} \sum_{j=1}^{m+1} \mathbf{v}^{(j)}_\xi \mathbf{v}^{(j)}_\xi + o_p(1) \tag{11}
\]

and

\[
\| \arg \sup_{k \in K} \hat{l}_2(\mathbf{k}; u, u) \| = \left( \sup_{k \in K^{(1)}} Q^{(1)}_{\xi k,u} \cdot \sup_{k \in K^{(2)}} Q^{(2)}_{\xi k,u} \cdot \cdots \cdot \sup_{k \in K^{(m)}} Q^{(m)}_{\xi k,u} \right)^{1/2} = O_p(1). \tag{12}
\]

In addition, by defining \( \hat{k}^{(j)}_u \equiv \arg \sup_{k \in K^{(j)}} Q^{(j)}_{\xi k,u} \) and \( \hat{K}_u \equiv (\hat{k}^{(1)}_u, \hat{k}^{(2)}_u, ..., \hat{k}^{(m)}_u)^T \), we have

\[
\| \hat{\beta}_k^* , u - \hat{\beta}_k^* , u \| = O_p(1/n) \tag{13}
\]

and

\[
\| \hat{\beta}_k^* , u - \hat{\beta}_k^* \| = O_p(1/\sqrt{n}). \tag{14}
\]

From (13), (14), and Murphy and van der Vaart (2000), we obtain

\[
\hat{l}_2 \left( \arg \sup_{k \in K} \hat{l}_2(\mathbf{k}; u, u); u \right)
= \hat{l}_2(\hat{K}_u, u) + o_p(1)
\]

\[
= - \sum_{j=1}^{m} \sum_{k \in K^{(j)}} Q^{(j)}_{\xi k,u} + \sum_{j=1}^{m+1} \left( \hat{\beta}^{(j)}_{k,u} - \hat{\beta}^{(j)}_{k^*,u} \right)^T \left( \sum_{i \in D^{(j)}} \{ z_i - h(t, \hat{\beta}^{(j)}_{k^*,u}) \} - \frac{1}{2} \mathbf{v}^{(j)}_\xi \mathbf{v}^{(j)}_\xi \right) \tag{15}
\]

See Web Appendix C for the detail of this derivation. Finally, from (11) and (15), we can obtain the following theorem.

**Theorem 1.** Under condition (2), the asymptotic bias in (5) is given by

\[
\text{E}[\hat{l}_2(\mathbf{k}^*, \beta_0^*; u)]
= \sum_{j=1}^{m} \text{E} \left( \sup_{k \in K^{(j)}} Q^{(j)}_{\xi k,u} + Q^{(j)}_{\hat{k}^{(j)}_u} \right) \tag{16}
+ \sum_{j=1}^{m+1} \text{tr} \left( \mathbf{A}^{(j)}_{\xi} \left( \mathbf{A}^{(j)}_{\xi} + \xi D^{(j)} I_p \right)^{-1} \right).
\]

We can regard the first and second terms on the right side of (16) as the biases for the change-point parameters \( k \) and the regression parameters \( \beta \), respectively.

### 3.2 Explicit expression of asymptotic bias

In the AIC for regular statistical models, the penalty is 2 for each parameter, regardless of whether its true value is a constant or converges to any value. In other words, it does not matter which setting is considered. On the other hand, in the AIC for conventional change-point models,
the penalty depends on the setting. In particular, the evaluation of the first term on the right side of (16) depends on whether \( \beta^{\tau(j+1)} - \beta^{\tau(j)} \) is a constant vector or converges to \( O_p \).

Here, we deem the latter case more important and natural. If \( \beta^{\tau(j+1)} - \beta^{\tau(j)} \) is a constant vector even in asymptotics, a clear change is expected to exist. In such a setting, the first term of the information criterion, that is, the goodness-of-fit term, almost entirely determines the model selection result, and the bias evaluation of the second term is less important. For cases in which it cannot be determined at first sight whether there are change-points, we need a more accurate evaluation of the second term, and the assumption of \( \beta^{\tau(j+1)} - \beta^{\tau(j)} \to O_p \) reflects such a case. Even if the existence of change-points is suspected at first glance, their existence is not absolute as long as the data size is finite. The assumption that \( \beta^{\tau(j+1)} - \beta^{\tau(j)} \) is a constant vector leads to asymptotic approximations that are too biased toward the existence of changes. Therefore, we consider it more natural to assume that \( \beta^{\tau(j+1)} - \beta^{\tau(j)} \) converges to \( O_p \).

From the above discussion, as in Section 1.5 of Csörgő and Horváth (1997), when estimating the parameters by maximizing the regularized log-partial likelihood function, we assume the following condition:

\[
\beta^{\tau(j+1)} - \beta^{\tau(j)} = \Delta^{(j)} \sqrt{\alpha_n} \quad (j \in \{1, 2, \ldots, m\}),
\]

\( O(1) \neq \alpha_n = o(n) \),

where \( \Delta^{(j)} \) is a constant vector. Then, instead of (10), we obtain \( \hat{k} = k^* = O_p(\alpha_n/n) \) from the key equation (7), and it can be seen that \( K = \{ k \mid k \in K^*(j), j \in \{1, 2, \ldots, m\} \} \), where \( K^*(j) = \{ k \mid k - k^{\tau(j)} = O(\alpha_n/n) \} \), thus yielding (11), (12), and (15). Hence, letting \( s \) be a vector with finite values, we assume that \( k = k^* + \alpha_n s/n \). Let \( \{ W_s \}_{s \in \mathbb{R}} \) denote two-sided standard Brownian motion with \( E(W_s) = 0 \) and \( V(W_s) = |s| \). In addition, let \( V_{\tau_1}(\tau_1, \tau_2, \sigma_1, \sigma_2) \) denote Brownian motion extending to both sides with drift coefficients of \( \tau_1 \) and \( \tau_2 \) and diffusion coefficients of \( \sigma_1 \) and \( \sigma_2 \). Moreover, let

\[
V^{\tau(j)}_{\xi,s} = V_{\tau_1}(\tau_1, \tau_2, \sigma_1, \sigma_2) + \left\{ \frac{1}{2} \Delta^{(j)} (A^{\tau(j)}_\xi + \xi D^{\tau(j)} I_p) \Delta^{(j)} + \frac{1}{2} \Delta^{(j)} (A^{\tau(j)}_\xi + \xi D^{\tau(j)} I_p) \Delta^{(j)} \right\}.
\]

Then, the following holds:

\[
Q^{(j)}_{\xi,s} + \alpha_n s/n \xrightarrow{d} V^{\tau(j)}_{\xi,s}.
\]

For explicit expression of the asymptotic bias, we need to evaluate the expectations of \( \sup_{\xi \in \mathbb{R}} V^{\tau(j)}_{\xi,s} \) and \( V^{\tau(j)}_{\xi,s} \), where \( V^{\tau(j)}_{\xi,s} \) is a copy of \( V^{\tau(j)}_{\xi,s} \). By the same flow of proof as in Ninomiya (2015), the following theorem can be obtained, although the paper does not take into account terms with respect to \( \xi \) due to regularization and does not even include the Cox model as a target. See Web Appendix D for the detail of this derivation and the definition of \( C(A^{(j)\dagger}, A^{(j)\dagger}) \).

**Theorem 2.** Under the conditions in Theorem 1 and (17), the asymptotic bias in (5) is given by

\[
E[b^{\tau(j)}(\hat{\beta}, \hat{\tau})] = 2 \sum_{j=1}^{m} C(A^{\tau(j)}_\xi, A^{\tau(j)}_\xi + \xi D^{\tau(j)} I_p)
+ \sum_{j=1}^{m+1} \text{tr}[A^{\tau(j)}_\xi(A^{\tau(j)}_\xi + \xi D^{\tau(j)} I_p)^{-1}].
\]

This gives an information criterion as the bias-corrected maximum regularized log-partial likelihood; however, because the asymptotic bias in (18) contains unknown parameters, they are replaced by consistent estimators, as in the TIC and the generalized information criterion (GIC, Konishi & Kitagawa, 1996). Specifically, as the consistent estimators of \( D^{\tau(j)} \) and \( A^{\tau(j)}_\xi \), we use \( \hat{D}^{\tau(j)} \equiv |\hat{D}^{\tau(j)}|/n \) and

\[
\hat{A}^{\tau(j)}_\xi(\hat{\beta}, \hat{\tau}) \equiv \frac{1}{n} \left[ \sum_{j \in \mathbb{D}^{(j)}} \{ z_i - h(t_i, \hat{\beta}^{(j)}) - \xi \hat{\beta}^{(j)} \} \right] + \left[ \sum_{j \in \mathbb{D}^{(j)}} \{ z_i - h(t_i, \hat{\beta}^{(j)}) - \xi \hat{\beta}^{(j)} \} \right]^T,
\]

respectively, where \( \hat{D}^{(j)} \equiv D(|\hat{k}^{(j-1)} - \hat{k}^{(j)}|) \). As a result, for the case where estimation is based on the partial likelihood with the addition of a regularization term in the \( L_2 \) norm, we propose the following information criterion for the Cox model with change-points:

\[
\text{AIC}_\xi = -2l(\hat{\beta}, \hat{\tau}; t) + 4 \sum_{j=1}^{m} \hat{C}(\hat{A}^{\tau(j)}_\xi(\hat{\beta}, \hat{\tau}), \hat{A}^{\tau(j)}_\xi(\hat{\beta}, \hat{k}))
+ \xi \hat{D}^{\tau(j)} I_p \} + 2 \sum_{j=1}^{m+1} \text{tr}[\hat{A}^{\tau(j)}_\xi(\hat{\beta}, \hat{\tau})(\hat{A}^{\tau(j)}_\xi(\hat{\beta}, \hat{k}) + \xi \hat{D}^{\tau(j)} I_p)^{-1}].
\]

where \( \hat{C}(A^{\tau(j)\dagger}, A^{\tau(j)\dagger}) \) is \( C(A^{\tau(j)\dagger}, A^{\tau(j)\dagger}) \) with \( \hat{\beta}^{\tau(j)}_\xi \) replaced by \( \hat{\beta} \). As the optimal value of \( \xi \), we have only to use the minimizer of this \( \text{AIC}_\xi \) with respect to \( \xi \).
Although so far we have discussed the model given by (1), in which all the regression parameters are structurally changed at change-points, even in the change-point model that some of the parameters are structurally changed as follows:

\[
\lambda(t \mid z) = \lambda_0(t) \exp(\beta_1^{(j)}z_1 + \beta_2^{(j)}z_2),
\]

the asymptotic bias is derived similarly to that given in (18) under the same conditions, and we can obtain the same AIC as that given in (19). Also, when the regularization parameter \( \xi \) is 0, \( \beta_2^{(j)} \) becomes equal to \( \beta^* \); then, by using \( C(A_{\xi}^{*}, A_{\xi}^{*(j)}) = \hat{C}[\hat{A}^{*}_{\xi}(\hat{\beta}, \hat{k}), \hat{A}^{*(j)}_{\xi}(\hat{\beta}, \hat{k})] = 3/2 \), we obtain the following corollary.

**Corollary 1.** Under the conditions in Theorem 1 and (17), the asymptotic bias in (5) and based on the conventional partial likelihood, which uses (3) with \( \xi = 0 \), is given by

\[
E[b(k^*, \beta^*)] = 3m + p(m + 1).
\]

From this, we can see that the asymptotic bias due to the change-point parameter is three times greater than that due to the regression parameter, which is consistent with the result in Ninomiya (2015). As a result, we propose the following criterion:

\[
\text{AIC} = -2\lambda(\tilde{\beta}, \tilde{k}; t) + 6m + 2p(m + 1),
\]  
(20)

which we call the AIC for the Cox model with change-points, for estimation based on the conventional partial likelihood method.

### 4 NUMERICAL EXPERIMENTS

In this section, we use the results of numerical experiments to examine the performance of the proposed AIC given in (20) (hereafter referred to simply as “AIC”) as an information criterion for estimation based on the conventional partial likelihood method without regularization. For comparison, we also consider the following information criterion:

\[
\text{AIC}_{\text{naive}} = -2\lambda(\tilde{\beta}, \tilde{k}; t) + 2m + 2p(m + 1),
\]

which handles the bias due to the change-point parameter in the same way that it handles the bias due to the regression parameter. To address the simplest setting, we assume that

\[
\lambda(t \mid z) = \begin{cases} 
\lambda_0(t) \exp(\beta_1^{(1)}z), & t \in [0, k) \\
\lambda_0(t) \exp(\beta_2^{(2)}z), & t \in [k, T)
\end{cases}
\]  
(21)

gives a univariate Cox model with one change-point \( k \), where \( z \) is a Bernoulli covariate with success probability of 1/2. As this experimental model has one change-point parameter and two regression parameters, the asymptotic bias evaluations for AIC and AIC_{naive} are \( 3 \times 1 + 1 \times 2 = 5 \) and \( 1 \times 1 + 1 \times 2 = 3 \), respectively.

First, to examine whether these penalty terms provide accurate approximations of the bias in the maximum log-partial likelihood, we numerically evaluated the bias with different true parameter values, different censoring rates and different event sizes in the model given by (21), where the censoring is assumed to be independent of covariates and survival time here and hereafter. The results are listed in Table 1. In every setting, the value was around 5, and a value of at least 5 was a more accurate approximation of the bias than 3, which indicate that AIC is a more accurate approximation of the Kullback–Leibler divergence than AIC_{naive}.

Second, to get a clear picture of how AIC makes model selections, we will compare it only with AIC_{naive} in the simplest setting. Specifically, we considered models given by

\[
\lambda(t \mid z) = \begin{cases} 
\lambda_0(t) \exp(\beta_1^{(j)}z), & t \in [0, k) \\
\lambda_0(t) \exp(\beta_2^{(j)}z), & t \in [k, T)
\end{cases}
\]  
(22)

with \( m = 0, m = 1, m = 2, \) and \( m = 3, \) where \( k^{(0)} = 0 \) and \( k^{(m+1)} = T \). Then, we selected the optimal model for each criterion. Table 2 summarizes the Kullback–Leibler divergence between the true and estimated distributions together with the selection probabilities under the true structure determined under a setting in which the number of change-points, \( m^* \), was 0 or 1. For the case of no change-points, that is, \( m^* = 0 \), we can see that, regardless of the event size, AIC could select the model with no change-point (i.e., with \( m = 0 \)) with a high probability of approximately 90% or higher. On the other hand, AIC_{naive} selected the model with change-points (i.e., the model with \( m > 0 \)) with a probability of approximately 50%. This result implies that AIC_{naive} underestimates the asymptotic bias and causes overfitting. Moreover, for the case of one change-point, that is, \( m^* = 1 \), when the event size and the amount of change were smaller, AIC was less likely than AIC_{naive} to select the number of true change-points. However, AIC gave a clearly smaller Kullback–Leibler divergence than AIC_{naive} under any setting, and we can
### Table 1 Bias in the maximum log-partial likelihood under a true model with one change-point.

| $\alpha^*$ | $\exp(\beta^{(2)}_*)$ | $r^*$ | $|D|:50$ | $|D|:100$ | $|D|:150$ | $|D|:200$ |
|---|---|---|---|---|---|---|
| 0.3 | 0.9 | 0.0 | 4.87 (0.30) | 4.97 (0.29) | 4.94 (0.28) | 4.86 (0.30) |
| | | 0.1 | 5.04 (0.36) | 5.42 (0.31) | 4.96 (0.33) | 4.55 (0.39) |
| | | 0.2 | 5.01 (0.34) | 5.03 (0.40) | 5.30 (0.46) | 5.29 (0.48) |
| 0.8 | 0.0 | 0.1 | 4.97 (0.33) | 4.93 (0.30) | 5.03 (0.31) | 4.85 (0.32) |
| | | 0.2 | 4.93 (0.34) | 5.05 (0.39) | 5.37 (0.48) | 5.75 (0.52) |
| 0.7 | 0.0 | 0.1 | 5.11 (0.36) | 5.13 (0.33) | 5.03 (0.35) | 4.73 (0.36) |
| | | 0.2 | 5.15 (0.39) | 5.27 (0.33) | 5.14 (0.41) | 4.61 (0.45) |
| 0.6 | 0.0 | 0.1 | 5.36 (0.42) | 5.29 (0.38) | 4.96 (0.40) | 4.95 (0.46) |
| | | 0.2 | 5.01 (0.34) | 5.03 (0.40) | 5.30 (0.46) | 5.29 (0.48) |
| 0.5 | 0.9 | 0.0 | 5.15 (0.37) | 4.99 (0.28) | 5.20 (0.27) | 5.02 (0.30) |
| | | 0.1 | 5.19 (0.37) | 5.21 (0.29) | 5.01 (0.35) | 4.57 (0.37) |
| | | 0.2 | 4.77 (0.35) | 5.01 (0.39) | 5.44 (0.46) | 5.26 (0.47) |
| 0.8 | 0.0 | 0.1 | 5.33 (0.39) | 4.91 (0.30) | 5.14 (0.28) | 5.06 (0.31) |
| | | 0.2 | 6.17 (1.26) | 5.24 (0.31) | 5.18 (0.37) | 4.72 (0.41) |
| 0.7 | 0.0 | 0.1 | 5.40 (0.43) | 5.01 (0.33) | 5.01 (0.31) | 4.85 (0.37) |
| | | 0.2 | 6.43 (1.33) | 5.39 (0.35) | 5.18 (0.39) | 4.86 (0.44) |
| 0.6 | 0.0 | 0.1 | 5.40 (0.43) | 4.84 (0.35) | 5.08 (0.34) | 4.98 (0.41) |
| | | 0.2 | 5.34 (0.41) | 5.46 (0.38) | 5.14 (0.48) | 4.83 (0.46) |
| 0.5 | 0.9 | 0.0 | 5.15 (0.37) | 4.99 (0.28) | 5.20 (0.27) | 5.02 (0.30) |
| | | 0.1 | 5.19 (0.37) | 5.21 (0.29) | 5.01 (0.35) | 4.57 (0.37) |
| | | 0.2 | 4.77 (0.35) | 5.01 (0.39) | 5.44 (0.46) | 5.26 (0.47) |
| 0.8 | 0.0 | 0.1 | 5.33 (0.39) | 4.91 (0.30) | 5.14 (0.28) | 5.06 (0.31) |
| | | 0.2 | 6.17 (1.26) | 5.24 (0.31) | 5.18 (0.37) | 4.72 (0.41) |
| 0.7 | 0.0 | 0.1 | 5.40 (0.43) | 5.01 (0.33) | 5.01 (0.31) | 4.85 (0.37) |
| | | 0.2 | 6.43 (1.33) | 5.39 (0.35) | 5.18 (0.39) | 4.86 (0.44) |
| 0.6 | 0.0 | 0.1 | 5.40 (0.43) | 4.84 (0.35) | 5.08 (0.34) | 4.98 (0.41) |
| | | 0.2 | 5.34 (0.41) | 5.46 (0.38) | 5.14 (0.48) | 4.83 (0.46) |

**Note:** The values are means (standard errors in parentheses) obtained by a Monte Carlo method through 100 iterations based on the model given by (21). The true change-point is $k^*$ satisfying $P(t < k^*) = \alpha^*$, the true regression parameters are $\beta^{(1)*} = 0$ and $\beta^{(2)*}$, the censoring rate is $r^*$, and the baseline hazard function is $\lambda_0(t) = 0.1$. The observed event size is $|D|$.

Thus say that AIC clearly selects the better model in terms of prediction.

Third, we compared the performance in a more practical setting. In the model given by (22) with $m = 0$, $m = 1$, $m = 2$, $m = 3$, and $m = 4$, we assumed that the number of true change-points was 2, that is, $m^* = 2$, and that the true change-points and the true amounts of changes were given randomly. For comparison, as well as AIC, we consider AIC_{naive} and the sequential test in He et al. (2013), where $|D|$ is the observed event size. Table 3 summarizes the Kullback–Leibler divergence between the true and selected models and the probability of selecting each model. Since AIC gave a smaller Kullback–Leibler divergence than AIC_{naive}, and the sequential test tended to select too few change-points, and this is one reason why their Kullback–Leibler divergence values were large.

Finally, to evaluate the performance of AIC_{s} in (19), we considered models given by

$$
\lambda(t \mid z) = \lambda_0(t) \exp(\beta_1^{(j)} z_1 + \beta_2^{(j)} z_2), \quad t \in [k^{(j-1)}, k^{(j)}), j \in \{1, 2, \ldots, m + 1\},
$$

with $m = 0$, $m = 1$, $m = 2$, $m = 3$, and $m = 4$. In view of the advantage of ridge-type regularization against multicollinearity, a positive correlation between the two Bernoulli covariates $z_1$ and $z_2$ was supposed. We assumed that the number of true change-points was 2 and that the true change-points and the true amounts of changes were given randomly. For comparison, as well as AIC, we consider AIC_{naive} and BIC_{naive}. Table 4 summarizes the Kullback–Leibler divergence and the selection probabilities. While AIC sometimes selected $m = 2$ with a higher probability than AIC_{s} in these experiments, AIC_{s} gave the
### Table 2

Kullback–Leibler divergence (K–L) between the true and estimated distributions, and the probability of selecting 0, 1, 2, or 3 change-points (%) under a true model with no or one change-point.

| $\alpha^*$ | $|D|$ | $\exp(\beta^{(2)})$ | $m^*$ | K–L | 0 (%) | 1 (%) | 2 (%) | 3 (%) |
|------------|------|---------------------|------|-----|-------|-------|-------|-------|
| 0.3        | 50   | 1.00                | 0    | AIC$_{naive}$ | 2.36  | 63    | 25    | 9     | 3     |
|            |      |                     |      | AIC    | 0.74  | 96    | 4     | 0     | 0     |
| 0.50       | 1    |                     | 0    | AIC$_{naive}$ | 3.51  | 45    | 31    | 20    | 4     |
|            |      |                     |      | AIC    | 2.04  | 80    | 20    | 0     | 0     |
| 0.25       | 1    |                     | 0    | AIC$_{naive}$ | 6.74  | 16    | 63    | 14    | 7     |
|            |      |                     |      | AIC    | 4.12  | 51    | 49    | 0     | 0     |
| 100        | 1.00 | 0                   | AIC$_{naive}$ | 4.97  | 37    | 11    | 30    | 22    |
|            |      |                     |      | AIC    | 0.96  | 95    | 2     | 2     | 1     |
| 0.50       | 1    |                     | 0    | AIC$_{naive}$ | 5.53  | 25    | 24    | 26    | 25    |
|            |      |                     |      | AIC    | 2.66  | 74    | 21    | 4     | 1     |
| 0.25       | 1    |                     | 0    | AIC$_{naive}$ | 5.10  | 5     | 56    | 20    | 19    |
|            |      |                     |      | AIC    | 4.14  | 25    | 68    | 5     | 2     |
| 0.5        | 50   | 1.00                | 0    | AIC$_{naive}$ | 3.25  | 57    | 19    | 20    | 4     |
|            |      |                     |      | AIC    | 1.17  | 93    | 5     | 2     | 0     |
| 0.50       | 1    |                     | 0    | AIC$_{naive}$ | 3.28  | 46    | 36    | 11    | 7     |
|            |      |                     |      | AIC    | 2.14  | 83    | 15    | 2     | 0     |
| 0.25       | 1    |                     | 0    | AIC$_{naive}$ | 5.30  | 11    | 64    | 19    | 6     |
|            |      |                     |      | AIC    | 3.26  | 44    | 54    | 2     | 0     |
| 100        | 1.00 | 0                   | AIC$_{naive}$ | 4.53  | 44    | 10    | 27    | 19    |
|            |      |                     |      | AIC    | 1.06  | 94    | 2     | 4     | 0     |
| 0.50       | 1    |                     | 0    | AIC$_{naive}$ | 5.59  | 17    | 30    | 29    | 24    |
|            |      |                     |      | AIC    | 3.19  | 66    | 24    | 9     | 1     |
| 0.25       | 1    |                     | 0    | AIC$_{naive}$ | 6.00  | 1     | 38    | 33    | 28    |
|            |      |                     |      | AIC    | 4.10  | 13    | 74    | 13    | 0     |

Note: The values are obtained by a Monte Carlo method through 100 iterations based on the model given by (22). The true structure is given by (21), where the true change-point is $k^*$ satisfying $P(t < k^*) = \alpha^*$, the true regression parameters are $\beta^{(1)*} = 0$ and $\beta^{(2)*}$, the censoring rate follows a discrete uniform distribution on $\{0.0, 0.1, 0.2\}$, and the baseline hazard function is $\lambda_0(t) = 0.1$. Because of this true structure, the number of the true change-points, $m^*$, is 0 or 1. The observed event size is $|D|$.

Abbreviation: AIC, Akaike information criterion.

The smallest Kullback–Leibler divergence than AIC, AIC$_{naive}$, and BIC$_{naive}$. Therefore, we can say that AIC$_{\xi}$ is the best information criterion in terms of prediction.

## 5 REAL DATA ANALYSIS

In this section, we apply the AIC in (20), AIC$_{naive}$, and BIC$_{naive}$ to data from a randomized, placebo-controlled clinical trial of patients with malignant glioma. One objective of this clinical trial was to examine whether implantation of carmustine-impregnated polymers into a brain tumor site after surgical resection of recurrent tumors could provide longer survival. A total of 222 patients were enrolled from 27 institutions: 110 patients were randomly assigned to the test group, while the other 112 were assigned to the control group. The details of clinical trial design and analysis results were reported in Brem et al. (1995). We inferred that the survival curves of the two groups, categorized by whether 75% or more of the tumor was resected in the clinical trial, would diverge after a certain period after resection.

Then, we searched for change-points by applying AIC, AIC$_{naive}$, and BIC$_{naive}$ to the data created by weighting each individual by 2 to check the behavior under a certain number of events. Specifically, letting $z$ be the variable that indicates whether 75% or more of the tumor is resected, we applied the three information criteria to select the optimal model among the models given by (22) with $m = 0$, $m = 1$, $m = 2$, $m = 3$, and $m = 4$. The results are listed in Table 5. Whereas AIC$_{naive}$ selected a model with four change-points, at 10.6, 12.0, 30.7, and 33.1 weeks, and whereas BIC$_{naive}$ selected a model with no change-points, AIC selected a model with one change-point, at 14.4 weeks. This suggests
that the three information criteria can produce considerably different results. Figure 1 shows Kaplan–Meier curves for the <75% and >75% resection groups. The two curves overlapped for less than 16 weeks, and then the difference between the curves increased, which makes it reasonable to expect that one structural change occurred around that time.

6 | EXTENSION

In this section, under the setting of $\xi = 0$, we extend the AIC in (20) to allow for model misspecification. As before, we defined $\hat{\beta}^*$ as the solution to the maximization problem $\arg\sup_{\beta} E(l(\beta, k^*; t))$. For the case without model misspecification, it denotes the true value of the regression parameter vector $\beta = (\beta^{(1)}T, \beta^{(2)}T, \ldots, \beta^{(m+1)}T)^T$ for the model given by (1). However, this model potentially assumes log-linearity for the relationship between the covariate $z$ and the hazard function $\lambda(t \mid z)$. Accordingly, application of this model to a situation in which this assumption does not hold would cause the model to be misspecified. Moreover, the model is misspecified in two cases: when the covariates that can be included in it are restricted, and when the conditional independence, given $z$, between the occurrence times of events and censoring, $y_1$ and $y_2$, does not hold. Hence, we derive an information criterion for the model given by (1) for the case of model misspecification. Note that $\hat{\beta}^*$ is not necessarily the true value for the case of model misspecification.

We incorporate the results of Struthers and Kalbfleisch (1986) and Lin and Wei (1989) for model misspecification into the way developed in Section 3.1. Then, we obtain the following corollary. See Web Appendix E for its derivation and the definitions of $A^{(j)}_0$ and $B^{(j)}_0$. 

---

**Table 3** Kullback–Leibler divergence (K–L) between the true and estimated distributions, and the probability of selecting 0, 1, 2, 3, or 4 change-points (%) under a true model with two change-points.

| $|D|$ | $\psi^*$ | K–L | 0 (%) | 1 (%) | 2 (%) | 3 (%) | 4 (%) |
|-----|---------|------|-------|-------|-------|-------|-------|
| 200 | 2.0     | AIC$_{naive}$ | 7.84  | 0     | 4     | 41    | 31    | 24    |
|     |         | AIC    | 6.30  | 2     | 26    | 68    | 4     | 0     |
|     |         | BIC$_{naive}$ | 6.82  | 7     | 40    | 51    | 2     | 0     |
|     |         | Sequential test | 7.03  | 11    | 38    | 50    | 1     | 0     |
| 1.8 |         | AIC$_{naive}$ | 8.12  | 0     | 5     | 40    | 28    | 27    |
|     |         | AIC    | 6.34  | 4     | 29    | 61    | 6     | 0     |
|     |         | BIC$_{naive}$ | 7.16  | 10    | 44    | 44    | 2     | 0     |
|     |         | Sequential test | 7.32  | 14    | 45    | 40    | 1     | 0     |
| 1.6 |         | AIC$_{naive}$ | 8.26  | 0     | 4     | 37    | 27    | 32    |
|     |         | AIC    | 6.36  | 5     | 35    | 49    | 9     | 2     |
|     |         | BIC$_{naive}$ | 6.81  | 11    | 53    | 33    | 3     | 0     |
|     |         | Sequential test | 6.65  | 13    | 53    | 34    | 0     | 0     |

| 400 | 1.4     | AIC$_{naive}$ | 8.98  | 0     | 1     | 26    | 33    | 40    |
|     |         | AIC    | 6.41  | 1     | 14    | 71    | 13    | 1     |
|     |         | BIC$_{naive}$ | 7.07  | 8     | 27    | 64    | 1     | 0     |
|     |         | Sequential test | 7.03  | 10    | 27    | 63    | 0     | 0     |
| 1.2 |         | AIC$_{naive}$ | 9.49  | 0     | 2     | 20    | 38    | 40    |
|     |         | AIC    | 6.91  | 5     | 23    | 61    | 10    | 1     |
|     |         | BIC$_{naive}$ | 7.37  | 10    | 40    | 49    | 1     | 0     |
|     |         | Sequential test | 7.04  | 13    | 34    | 53    | 0     | 0     |
| 1.0 |         | AIC$_{naive}$ | 9.29  | 1     | 4     | 22    | 35    | 38    |
|     |         | AIC    | 6.66  | 3     | 30    | 56    | 9     | 2     |
|     |         | BIC$_{naive}$ | 7.33  | 14    | 49    | 36    | 1     | 0     |
|     |         | Sequential test | 7.15  | 16    | 45    | 38    | 1     | 0     |

Note: The values are obtained by a Monte Carlo method through 100 iterations with varying true parameters based on the model given by (22). The observed event size is $|D|$. The true change-points are $k^{(1)}$ and $k^{(2)}$ satisfying $P(t < k^{(1)}) = \alpha_1^*$ and $P(k^{(1)} < t < k^{(2)}) = \alpha_2^*$, where $\alpha_1^*$ and $\alpha_2^*$ follow a continuous uniform distribution over [0.1, 0.4]. The true amounts of changes are $\exp(\beta^{(2)}_1)/\exp(\beta^{(2)}_1) = 2u_1^*(\psi + v_1^*)$ and $\exp(\beta^{(3)}_1)/\exp(\beta^{(2)}_1) = 2u_2^*(\psi + v_2^*)$, where $u_1^*$ and $u_2^*$ follow a discrete uniform distribution on {-1, 1}, $v_1^*$ and $v_2^*$ follow a continuous uniform distribution over [0, 1], and $\beta^{(2)}_1 = 0$. The censoring rate follows a discrete uniform distribution on {0.0, 0.1, 0.2}, and the baseline hazard function is $\lambda_0(t) = 0.1$.

Abbreviations: AIC, Akaike information criterion; BIC, Bayesian information criterion.
TABLE 4 Kullback–Leibler divergence (K–L) between the true and estimated distributions, and the probability of selecting 0, 1, 2, 3, or 4 change-points (%) under a true model with two change-points in the presence of correlations between covariates.

| | | | K–L | 0 (%) | 1 (%) | 2 (%) | 3 (%) | 4 (%) |
|---|---|---|---|---|---|---|---|---|
| | | AIC$_{\text{naive}}$ | 9.81 | 0 | 12 | 40 | 32 | 16 |
| | | AIC | 7.75 | 1 | 39 | 52 | 8 | 0 |
| | | AIC$_{\xi}$ | 6.64 | 0 | 34 | 46 | 16 | 4 |
| | | BIC$_{\text{naive}}$ | 8.80 | 18 | 62 | 20 | 0 | 0 |
| 0.8 | | AIC$_{\text{naive}}$ | 7.90 | 0 | 30 | 60 | 8 | 2 |
| | | AIC | 7.47 | 2 | 57 | 41 | 0 | 0 |
| | | AIC$_{\xi}$ | 6.52 | 0 | 48 | 49 | 3 | 0 |
| | | BIC$_{\text{naive}}$ | 8.20 | 11 | 67 | 22 | 0 | 0 |
| 1.6 | | AIC$_{\text{naive}}$ | 9.96 | 0 | 11 | 35 | 34 | 20 |
| | | AIC | 7.74 | 3 | 49 | 39 | 9 | 0 |
| | | AIC$_{\xi}$ | 6.24 | 1 | 34 | 43 | 19 | 3 |
| | | BIC$_{\text{naive}}$ | 8.80 | 18 | 62 | 20 | 0 | 0 |
| 400 | 1.4 | AIC$_{\text{naive}}$ | 12.11 | 0 | 3 | 32 | 30 | 35 |
| | | AIC | 8.88 | 0 | 23 | 59 | 18 | 0 |
| | | AIC$_{\xi}$ | 7.73 | 0 | 20 | 55 | 20 | 5 |
| | | BIC$_{\text{naive}}$ | 9.13 | 8 | 54 | 38 | 0 | 0 |
| 0.8 | | AIC$_{\text{naive}}$ | 7.89 | 0 | 14 | 45 | 25 | 16 |
| | | AIC | 7.64 | 0 | 35 | 56 | 9 | 0 |
| | | AIC$_{\xi}$ | 6.32 | 0 | 29 | 56 | 13 | 2 |
| | | BIC$_{\text{naive}}$ | 7.81 | 7 | 60 | 33 | 0 | 0 |
| 1.0 | | AIC$_{\text{naive}}$ | 12.33 | 0 | 9 | 25 | 32 | 34 |
| | | AIC | 8.36 | 1 | 32 | 55 | 11 | 1 |
| | | AIC$_{\xi}$ | 7.16 | 0 | 26 | 48 | 17 | 9 |
| | | BIC$_{\text{naive}}$ | 8.85 | 17 | 66 | 17 | 0 | 0 |
| 0.8 | | AIC$_{\text{naive}}$ | 9.56 | 0 | 21 | 41 | 20 | 18 |
| | | AIC | 7.32 | 1 | 47 | 46 | 5 | 1 |
| | | AIC$_{\xi}$ | 6.23 | 0 | 38 | 44 | 16 | 2 |
| | | BIC$_{\text{naive}}$ | 8.08 | 15 | 60 | 25 | 0 | 0 |

Note: The values are obtained by a Monte Carlo method through 100 iterations with varying true parameters based on the model given by (23). The observed event size is $|D|$. The true change-points are $k^{(1)}$ and $k^{(2)}$ satisfying $P(t < k^{(1)}) = \alpha_1^*$ and $P(k^{(1)} < t < k^{(2)}) = \alpha_2^*$, where $\alpha_1^*$ and $\alpha_2^*$ follow a continuous uniform distribution over $[0.1, 0.4]$. The true amounts of changes are $\exp(\beta^{(2)})/\exp(\beta^{(1)}) = 2u^*_1(\phi^*_1 + v^*_1)$ and $\exp(\beta^{(3)})/\exp(\beta^{(2)}) = 2u^*_2(\phi^*_2 + v^*_2)$, where $u^*_1$ and $u^*_2$ follow a discrete uniform distribution on $\{-1, 1\}$, $v^*_1$ and $v^*_2$ follow a continuous uniform distribution over $[0, 1]$, and $\beta^{(1)} = 0$. The correlation between $z_1$ and $z_2$ is $\rho^*$, which indicates the size of multicollinearity. The censoring rate follows a discrete uniform distribution on $\{0.0, 0.1, 0.2\}$, and the baseline hazard function is $\lambda_0(t) = 0.1$.

Abbreviations: AIC, Akaike information criterion; BIC, Bayesian information criterion.

TABLE 5 Change-point estimates $\hat{k}$, maximum log-partial likelihood $l(\hat{\theta}, \hat{k}; t)$, AIC$_{\text{naive}}$, AIC, and BIC$_{\text{naive}}$ obtained from real clinical trial data.

| | $k^{(1)}$ | $k^{(2)}$ | $k^{(3)}$ | $l(\hat{\theta}, \hat{k}; t)$ | AIC$_{\text{naive}}$ | AIC | BIC$_{\text{naive}}$ |
|---|---|---|---|---|---|---|---|
| 0 | -2169.65 | 4341.29 | 4341.29 | 4345.32 |
| 1 | -2164.92 | 4335.83 | 4339.83 | 4347.91 |
| 2 | -2161.69 | 4333.39 | 4341.39 | 4353.52 |
| 3 | -2158.72 | 4331.45 | 4343.45 | 4359.63 |
| 4 | -2156.36 | 4330.72 | 4346.72 | 4366.95 |

Abbreviations: AIC, Akaike information criterion; BIC, Bayesian information criterion.
FIGURE 1 Kaplan–Meier curves for the <75% and >75% resection groups. The upper limit for the change-point is set at 48 weeks because approximately 95% of events in the group with <75% resection occurred by 48 weeks, with the remaining events occurring after approximately 80 weeks.

Corollary 2. Under the condition in Theorem 1, even if model misspecification with \( \xi = 0 \) in (3) exists, the asymptotic bias in (5) is given by

\[
E[b(k^*, \beta^*)] = \sum_{j=1}^{m} \left( \sup_{k \in K(j)} Q_{k,t}^{(j)} + \sum_{j=1}^{m+1} \text{tr}(A_0^{s(j)} B_0^{s(j)-1}) \right)
\]

Let us assume the following condition:

\[
\beta_{s(j+1)} - \beta_{s(j)} = \Delta_{p,j} / \sqrt{\alpha_n} \quad (j \in \{1, 2, ..., m\}), \\
O(1) \neq \alpha_n = o(n). \tag{24}
\]

As in Section 3.2, the asymptotic behavior of the change-point estimator can be investigated under this condition. Then, we obtain the following corollary. See Web Appendix F for its derivation.

Corollary 3. Under the conditions in Theorem 2 and (24), even if model misspecification with \( \xi = 0 \) in (3) exists, the asymptotic bias in (5) is given by

\[
E[b(k^*, \beta^*)] = \sum_{j=1}^{m} C(A_0^{s(j)}, B_0^{s(j)} + \sum_{j=1}^{m+1} \text{tr}(A_0^{s(j)} B_0^{s(j)-1}).
\]

While this gives an information criterion via the bias-corrected maximum log-partial likelihood, because the asymptotic bias in (3) contains unknown parameters, they are replaced by consistent estimators, as in (19). As a result, we propose the following information criterion for the Cox model with change-points in cases of model misspecification:

\[
\text{TIC} = -2\ell(\hat{\beta}, \hat{k}; t) + 4 \sum_{j=1}^{m} \hat{C}(A_0^{s(j)} B_0^{s(j)}) \\
+ 2 \sum_{j=1}^{m+1} \text{tr}(A_0^{s(j)} B_0^{s(j)-1}),
\]

where

\[
\hat{A}_0^{s(j)} (\hat{\beta}, \hat{k}) \equiv \frac{1}{n} w^{s(j)} (\hat{\beta}, \hat{k}) w^{s(j)} (\hat{\beta}, \hat{k})^T
\]

and

\[
\hat{B}_0^{s(j)} (\hat{\beta}, \hat{k}) \equiv \frac{1}{n} \sum_{i \in D^{s(j)}} \{H(t_i, \hat{\beta}, \hat{k}) - h(t_i, \hat{\beta}, \hat{k}) h(t_i, \hat{\beta}, \hat{k})^T\}.
\]

To evaluate the performance of TIC under the presence of model misspecification, we considered true structures given by

\[
\lambda(t | z) = \lambda_0(t) \exp(\beta_1 z_1 + \beta_2 z_2), \quad t \in [k^{(j-1)}, k^{(j)}], \\
j \in \{1, 2, ..., m+1\}, \tag{25}
\]

with \( m = 2 \), where \( z_1 \) and \( z_2 \) are uncorrelated two covariates, and \( \beta_2 \neq 0 \). On the other hand, the candidates to be selected are models given by (25) with \( m = 0, m = 1, m = 2, m = 3, \) and \( m = 4 \), where \( \beta_2 = 0 \). These are misspecified models because of the existence of \( z_2 \). Table 6 summarizes the Kullback–Leibler divergence between the true and selected models and the probability of selecting each model. In these experiments, TIC is slightly better than AIC in terms of prediction, while the selection probabilities for both criteria are almost the same.
### TABLE 6
Kullback–Leibler divergence (K–L) between the true and estimated distributions, and the probability of selecting 0, 1, 2, 3, or 4 change-points (%) under a true model with two change-points in the presence of model misspecification.

| $|D|$ | $\psi^*$ | K–L | 0 (%) | 1 (%) | 2 (%) | 3 (%) | 4 (%) |
|---|---|---|---|---|---|---|---|
| 200 | 2.0 | AIC$_{\text{naive}}$ | 23.82 | 0 | 5 | 30 | 31 | 34 |
| | | AIC | 21.43 | 0 | 34 | 59 | 6 | 1 |
| | | TIC | 21.22 | 0 | 29 | 63 | 7 | 1 |
| | | BIC$_{\text{naive}}$ | 21.89 | 9 | 49 | 40 | 4 | 0 |
| 1.8 | AIC$_{\text{naive}}$ | 23.80 | 0 | 8 | 28 | 29 | 35 |
| | | AIC | 21.35 | 2 | 33 | 58 | 6 | 1 |
| | | TIC | 21.29 | 2 | 31 | 60 | 6 | 1 |
| | | BIC$_{\text{naive}}$ | 21.62 | 12 | 47 | 38 | 3 | 0 |
| 1.6 | AIC$_{\text{naive}}$ | 24.11 | 0 | 9 | 23 | 28 | 40 |
| | | AIC | 21.46 | 9 | 42 | 42 | 6 | 1 |
| | | TIC | 21.31 | 9 | 38 | 45 | 7 | 1 |
| | | BIC$_{\text{naive}}$ | 21.53 | 16 | 53 | 29 | 2 | 0 |
| 400 | 1.4 | AIC$_{\text{naive}}$ | 38.29 | 0 | 1 | 26 | 30 | 43 |
| | | AIC | 36.19 | 1 | 19 | 67 | 11 | 2 |
| | | TIC | 36.07 | 1 | 19 | 68 | 10 | 1 |
| | | BIC$_{\text{naive}}$ | 36.58 | 7 | 43 | 49 | 1 | 0 |
| 1.2 | AIC$_{\text{naive}}$ | 38.75 | 0 | 1 | 22 | 35 | 42 |
| | | AIC | 36.69 | 4 | 23 | 56 | 14 | 3 |
| | | TIC | 36.69 | 4 | 22 | 56 | 15 | 3 |
| | | BIC$_{\text{naive}}$ | 36.80 | 10 | 48 | 41 | 1 | 0 |
| 1.0 | AIC$_{\text{naive}}$ | 38.66 | 0 | 6 | 18 | 32 | 44 |
| | | AIC | 36.22 | 4 | 35 | 46 | 11 | 4 |
| | | TIC | 36.19 | 4 | 35 | 46 | 12 | 3 |
| | | BIC$_{\text{naive}}$ | 36.34 | 13 | 63 | 23 | 1 | 0 |

Note: The values are obtained by a Monte Carlo method through 100 iterations with varying true parameters based on the model given by (25). The observed event size is $|D|$. The true change-points are $k^{(1)}$ and $k^{(2)}$ satisfying $P(t < k^{(1)}) = \alpha_1^*$ and $P(k^{(1)} < t < k^{(2)}) = \alpha_2^*$, where $\alpha_1^*$ and $\alpha_2^*$ follow a continuous uniform distribution over $[0.1, 0.4]$. The true amounts of changes are $\exp(\beta_2^{(1)})/\exp(\beta_1^{(1)}) = 2u_1^*$ $(\psi_1^* + \psi_2^*)$ and $\exp(\beta_2^{(3)})/\exp(\beta_1^{(2)}) = 2u_2^*$ $(\psi_1^* + \psi_2^*)$, where $u_1^*$ and $u_2^*$ follow a discrete uniform distribution on $\{-1, 1\}$, $\psi_1^*$ and $\psi_2^*$ follow a continuous uniform distribution over $[0, 1]$, and $\beta_2^{(1)} = 0$. The parameter generating the model misspecification, $\beta_2^{(1)}$, and the censoring rate follow discrete uniform distributions on $[0.5, 1.0]$ and $[0.0, 0.1, 0.2]$, respectively, and the baseline hazard function is $\lambda_0(t) = 0.1$.

Abbreviations: AIC, Akaike information criterion; BIC, Bayesian information criterion; TIC, Takeuchi information criterion.

Note that AIC clearly outperforms AIC$_{\text{naive}}$ even if models are misspecified.

### 7 CONCLUSION

In light of the high demand for change-point detection in survival time analysis, this paper has derived AIC-type information criteria for the Cox model with change-points. First, we evaluated the asymptotic bias of the regularized maximum log-partial likelihood, and we showed via Theorem 1 that the asymptotic bias caused by a change-point can be expressed in terms of a two-sided random walk. Then, by assuming an additional natural condition for asymptotics, we showed via Theorem 2 that it can be expressed in a simple form. When there is no regularization term, it can be more easily written as 3.

The model here has a different aspect from conventional change-point models in that the time and outcome variables are the same. This indicates the need for new asymptotics; we have however revealed that as long as the partial likelihood method is used, it is sufficient to deal with conventional asymptotics. Through numerical experiments, we demonstrated that the evaluated asymptotic bias could be approximated with high accuracy. Furthermore, regarding the original purpose of AIC-type information criteria, which is to give an estimate close to the true structure, the proposed information criteria are clearly superior to the formal information criteria. Moreover, through real data analysis, we indicated that the formal information criteria would easily lead to different results from the proposed AIC. Although this paper addressed to mitigate the proportional hazard property, the proposed AIC still relied on log-linearity. Accordingly,
we extended it to the TIC reasonable in terms of original derivation under model misspecification.

Although we addressed a model in which the hazard function changes with time (i.e., a change-point model for time), models in which the hazard function changes with covariate values (i.e., change-point models for covariates) have also been discussed, especially in recent years (e.g., Pons, 2003; Lee & Lam, 2020; and Wang et al., 2021). While a change-point model for time basically considers a jump model with abrupt changes, change-point models for covariates also often consider a model with gradual changes. In a jump-type model for covariates, as in a change-point model for time, the estimator of the change-point parameter has been reported to converge faster than that of the regression parameter. In contrast, in a gradual-change model for covariates, the estimator of the change-point parameter converges at the same rate as that of the regression parameter, and it has been reported to have asymptotic normality. For change-point detection, test-based methods using asymptotic normality have been proposed. However, information-criterion-based methods have not been proposed, and it will be necessary to develop them. Because the convergence speeds for the estimators are different in the two change-point models, which implies a difference in the accuracies of the estimators, we expect that the penalty terms of the information criteria for the two models will be considerably different. Specifically, the instances of two-sided Brownian motion appearing in the limit are expected to be different, which will necessitate a new evaluation of the expectation.

In survival time analysis, joint modeling, which simultaneously models repeatedly measured covariates and survival time data, has gained attention (see, e.g., Henderson et al., 2000). Because this approach is an extension of the Cox model, change-point analysis for both time and covariates in this model will be necessary. The first difficulty is in the use of the profile likelihood, which can be regarded as an extension of the partial likelihood. The asymptotic theory was constructed in Zeng and Cai (2005), and the problem will be to tune the theory and reconcile it with the theory used in this paper. Another difficulty in joint modeling is that construction of the information criterion itself is also a hurdle. Regarding this difficulty, because we will deal with a semiparametric model that is also for repeatedly measured covariates and usually includes a random-effects term, the construction of the AIC or conditional AIC, as in Xu et al. (2009) and Donohue et al. (2011), will not be trivial.

ACKNOWLEDGMENTS
The authors are grateful to the editors and referees for their helpful and constructive comments. This research was supported by a JSPS Grant-in-Aid for Scientific Research (16K00050).

DATA AVAILABILITY STATEMENT
The data supporting the findings of this paper are available from the website of Piantadosi (2017), https://www.wiley.com/en-us/Clinical+Trials%3A+A+%3BMethodologic+Perspective%2C+3rd+Edition-p-9781118959206.

ORCID
Ryoto Ozaki https://orcid.org/0000-0002-1142-4162
Yoshiyuki Ninomiya https://orcid.org/0000-0001-5158-9108

REFERENCES
Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N. & Csaki, F. (Eds.) 2nd International Symposium on Information Theory. Budapest: Akademiai Kiado, pp. 267–281.
Brem, H., Piantadosi, S., Burger, P.C., Walker, M., Selker, R., Vick, N.A. et al. (1995) Placebo-controlled trial of safety and efficacy of intraoperative controlled delivery by biodegradable polymers of chemotherapy for recurrent gliomas. Lancet, 345(8956), 1008–1012.
Cox, D.R. (1972) Regression models and life-tables. Journal of the Royal Statistical Society: Series B, 34(2), 187–202.
Csörgő, M. & Horváth, L. (1997) Limit theorems in change-point analysis. New York: Wiley.
Dacunha-Castelle, D. & Gassiat, E. (1999) Testing the order of a model using locally conic parametrization: population mixtures and stationary ARMA processes. Annals of Statistics, 27(4), 1178–1209.
Donohue, M., Overholser, R., Xu, R. & Vaida, F. (2011) Conditional Akaike information under generalized linear and proportional hazards mixed models. Biometrika, 98(3), 685–700.
Gill, R. & Schumacher, M. (1987) A simple test of the proportional hazards assumption. Biometrika, 74(2), 289–300.
He, P., Fang, L. & Su, Z. (2013) A sequential testing approach to detecting multiple change points in the proportional hazards model. Statistics in Medicine, 32(7), 1239–1245.
Henderson, R., Diggle, P. & Dobson, A. (2000) Joint modelling of longitudinal measurements and event time data. Biostatistics, 1(4), 465–480.
Hess, K.R. (1995) Graphical methods for assessing violations of the proportional hazards assumption in Cox regression. Statistics in Medicine, 14(15), 1707–1723.
Hoerl, A.E. & Kennard, R.W. (1970) Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55–67.
Konishi, S. & Kitagawa, G. (1996) Generalised information criteria in model selection. Biometrika, 83(4), 875–890.
Lee, C.Y. & Lam, K.F. (2020) Survival analysis with change-points in covariate effects. Statistical Methods in Medical Research, 29(11), 3235–3248.
Lin, D.Y. & Wei, L.-J. (1989) The robust inference for the Cox proportional hazards model. Journal of the American Statistical Association, 84(408), 1074–1078.

1992, 16(3), 294–301.

1996, 31(1), 87–94.

2005, 51(1), 124–133.

2000, 2(1), 87–96.

2011, 30(2), 242–258.

2004, 16(2), 161–171.

2001, 32(3), 93–103.
Liu, M., Lu, W. & Shao, Y. (2008) A Monte Carlo approach for change-point detection in the Cox proportional hazards model. *Statistics in Medicine*, 27(19), 3894–3909.

Murphy, S.A. & van der Vaart, A.W. (2000) On profile likelihood. *Journal of the American Statistical Association*, 95(450), 449–465.

Ninomiya, Y. (2015) Change-point model selection via AIC. *Annals of the Institute of Statistical Mathematics*, 67(5), 943–961.

Piantadosi, S. (2017) *Clinical trials: a methodologic perspective*. Wiley.

Pons, O. (2002) Estimation in a Cox regression model with a change-point at an unknown time. *Statistics*, 36(2), 101–124.

Pons, O. (2003) Estimation in a Cox regression model with a change-point according to a threshold in a covariate. *Annals of Statistics*, 31(2), 442–463.

Siegmund, D.O. (2004) Model selection in irregular problems: applications to mapping quantitative trait loci. *Biometrika*, 91(4), 785–800.

Struthers, C.A. & Kalbfleisch, J.D. (1986) Misspecified proportional hazard models. *Biometrika*, 73(2), 363–369.

Takeuchi, K. (1976) The distribution of information statistics and the criterion of goodness of fit of models. *Mathematical Science*, 153, 12–18 (in Japanese).

Tsiatis, A.A. (1981) A large sample study of Cox’s regression model. *Annals of Statistics*, 9(1), 93–108.

Uno, H., Claggett, B., Tian, L., Inoue, E., Gallo, P., Miyata, T., et al. (2014) Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *Journal of Clinical Oncology*, 32(22), 2380–2385.

van Houwelingen, H.C., Bruinsma, T., Hart, A.A., van’t Veer, L.J. & Wessels, L.F. (2006) Cross-validated Cox regression on microarray gene expression data. *Statistics in Medicine*, 25(18), 3201–3216.

Wang, B., Li, J. & Wang, X. (2021) Change point detection in Cox proportional hazards mixture cure model. *Statistical Methods in Medical Research*, 30(2), 440–457.

Witten, D.M. & Tibshirani, R. (2010) Survival analysis with high-dimensional covariates. *Statistical Methods in Medical Research*, 19(1), 29–51.

Xu, R., Vaida, F. & Harrington, D.P. (2009) Using profile likelihood for semiparametric model selection with application to proportional hazards mixed models. *Statistica Sinica*, 19(2), 819–842.

Zeng, D. & Cai, J. (2005) Asymptotic results for maximum likelihood estimators in joint analysis of repeated measurements and survival time. *Annals of Statistics*, 33(5), 2132–2163.

**SUPPORTING INFORMATION**

Web Appendices referenced in Sections 3 and 6 and the R code implementing the proposed criteria are available with this paper at the Biometrics website on Wiley Online Library.

How to cite this article: Ozaki, R. & Ninomiya, Y. (2023) Information criteria for detecting change-points in the Cox proportional hazards model. *Biometrics*, 79, 3050–3065. https://doi.org/10.1111/biom.13855