Statistics
in Medicine

Research Article

Received 21 April 2016, Accepted 19 May 2017 Published online 12 June 2017 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.7368

Information criteria for Firth's penalized partial likelihood approach in Cox regression models

Kengo Nagashima*† and Yasunori Sato

In the estimation of Cox regression models, maximum partial likelihood estimates might be infinite in a monotone likelihood setting, where partial likelihood converges to a finite value and parameter estimates converge to infinite values. To address monotone likelihood, previous studies have applied Firth's bias correction method to Cox regression models. However, while the model selection criteria for Firth's penalized partial likelihood approach have not yet been studied, a heuristic AIC-type information criterion can be used in a statistical package. Application of the heuristic information criterion to data obtained from a prospective observational study of patients with multiple brain metastases indicated that the heuristic information criterion selects models with many parameters and ignores the adequacy of the model. Moreover, we showed that the heuristic information criterion tends to select models with many regression parameters as the sample size increases. Thereby, in the present study, we propose an alternative AIC-type information criterion based on the risk function. A Bayesian information criterion type was also evaluated. Further, the presented simulation results confirm that the proposed criteria performed well in a monotone likelihood setting. The proposed AIC-type criterion was applied to prospective observational study data. © 2017 The Authors. *Statistics in Medicine* Published by John Wiley & Sons Ltd

Keywords: Akaike's information criterion; model selection; monotone likelihood; penalized partial likelihood; survival analysis

1. Introduction

The Cox regression model [1] is one of the most useful and widely used tools in survival analysis. In Cox regression model estimations, maximum partial likelihood estimates may be infinite in monotone likelihood situations [2]. In such cases, partial likelihood converges to a finite value, and the parameter estimates and standard errors converge to infinite values; hence, these results are not interpretable. Such problems arise, for example, in the presence of unbalanced covariates, large parameter effects, and/or heavy censoring.

To address this monotone likelihood problem, Heinze and Schemper proposed Firth's penalized partial likelihood approach [3]. They directly applied to the Cox regression model Firth's bias correction method [4], which aims to remove asymptotic bias from maximum likelihood estimates in exponential families with canonical link functions. Firth's penalized partial likelihood approach reduces asymptotic bias and addresses the monotone likelihood problem [3, 5]. Firth's bias correction method was also applied to logistic regression models to address the separation problem [5–7], which is similar to the monotone likelihood problem. This approach reduces asymptotic bias and also overcomes the separation problem.

In this study, we discuss the model selection criteria for Firth's penalized partial likelihood approach based on Akaike's information criterion (AIC) [8] and Bayesian information criterion (BIC) [9]. Although

Department of Global Clinical Research, Graduate School of Medicine, Chiba University, 1-8-1 Inohana, Chuo-ku, Chiba 260-8670, Japan

Correspondence to: Kengo Nagashima, Department of Global Clinical Research, Graduate School of Medicine, Chiba University, 1-8-1 Inohana, Chuo-ku, Chiba 260-8670, Japan.

†*E-mail: nshi@chiba-u.jp*

model selection is an important issue in data analysis, the model selection criteria for Firth's penalized partial likelihood approach have never been studied. To our best knowledge, only the SAS PHREG procedure can be used to obtain an AIC-type heuristic information criterion, AIC* = −2log(maximum penalized partial likelihood)+2p, where p is the number of regression parameters, and other major statistical software (e.g., Stata and R) can only output the log penalized partial likelihood. However, AIC* is not theoretically justified, and especially, we find that AIC* tends to select a model that has a large number of regression parameters as $n \to \infty$, where n is the sample size; that is, AIC* does not have the important property of avoiding over-fitting. This result indicates that AIC* is not a suitable model selection criterion. Similarily, the SAS PHREG procedure implements a BIC-type heuristic information criterion, BIC* = −2log(maximum penalized partial likelihood)+ log d, where d is the number of events. BIC* is not also theoretically justified. Therefore, we consider alternative model selection criteria in this setting.

The remainder of this paper is organized as follows. In Section 2, we introduce motivating data and issues of AIC*. Section 3 briefly reviews Firth's bias correction method and penalized partial likelihood approach, discusses the fundamental problems of AIC* and BIC*, and proposes appropriate information criteria. Section 4 presents the simulation results to demonstrate the performance of the criteria and to check the property of AIC* holds. Section 5 applies the proposed method to real data, and Section 6 concludes the paper with a brief discussion.

2. Motivating example

Yamamoto *et al.* [10] collected time-to-event (e.g., death, local recurrence, and leptomeningeal dissemination) data for 1194 cancer patients with multiple brain metastases. Secondary end points of this study include time to leptomeningeal dissemination (the data on 928 patients were censored, while the MRI results of 121 patients (10%), that is, those who suffered an early death or who deteriorated markedly soon after stereotactic radiosurgery, were not available; 145 patients had an event). We analyzed the following covariates: *age* (<65, ≥65), *sex* (female, male), *kps* (Karnofsky performance status; ≥80, ≤70), *ntumor* (number of tumors; 1, 2–4, 5–10), *diameter* (maximum diameter of largest tumor; <1.6 cm, ≥1.6 cm), *volume* (cumulative tumor volume; <1.9 mL, ≥1.9 mL), *ptumor* (primary tumor category; lung, breast, gastrointestinal, kidney, other), *status* (extracerebral disease status; not controlled, controlled), and *neuro* (neurological symptoms; no, yes). To analyze the competing risk end point, leptomeningeal dissemination, we used cause-specific proportional hazard models, which are identical to usual Cox regression models [11].

The descriptive statistics for the study data are shown in Table I. The data have heavy censoring, and the number of events differs considerably for the primary tumor categories. In particular, the kidney cancer group has no events. Further, as we illustrate below, monotone likelihood was observed in these data because of the primary tumor categories, while the parameter estimate of the kidney cancer group converges to $-\infty$.

Next, we consider the model selection based on AIC*, the results of which are shown in Table II. The full model, which includes all the covariates, was selected by AIC* as the best. In the best model, the hazard ratios were estimated by using Firth's penalized partial likelihood approach (Table III). To illustrate the problem of monotone likelihood, the hazard ratios estimated by using a usual Cox regression model are also shown. For the usual Cox regression model, when monotone likelihood occurred, the hazard ratio of the kidney cancer group was 0.00 (= exp($-\infty$)), standard error was 543.30, p-value of the Wald test was 0.98, and p-value of the likelihood ratio test was <0.01. Although the number of events for lung cancer was considerably larger than that for kidney cancer (Table I), a large p-value was observed in the Wald test. On the contrary, the results derived by using Firth's penalized partial likelihood approach were plausible. The hazard ratio was 0.12, and standard error was 1.43 (Table III); therefore, usual Cox regression models were unsatisfactory in the presence of monotone likelihood.

Now, we return to the model selection result based on AIC* when using Firth's penalized partial likelihood approach. As shown in Table II, model selection based on AIC* tends to select models that have many parameters, and to support this statement, we discuss the theoretical property of AIC* in Section 3.3. In Table III, the best model under AIC* includes variables that have considerably small effects such as *age* (*HR* = 1.00, p-value = 0.98) and *status* (*HR* = 1.03, p-value = 0.89), whose p-values were very large. Although these variables have little association with the time-to-event, such variables were included in the best model and subsequent models ranked in the top 5. Indeed, because model selection based on AIC* performs badly, we propose an alternative approach herein to address this problem.

Table I. Number of events (leptomeningeal dissemination) and censored values for the study data (n = 1073).

Covariate	Group	Event	Censored	% Censored
age	<65	69	393	85.1
	≥65	76	535	87.6
sex	Female	62	370	85.6
	Male	83	558	87.1
kps	≥80	132	810	86.0
	≤70	13	118	90.1
ntumor	1	49	365	88.2
	2–4	61	412	87.1
	5–10	35	151	81.2
diameter	<1.6 cm	76	455	85.7
	≥1.6 cm	69	473	87.3
volume	<1.9 mL	75	459	86.0
	≥1.9 mL	70	469	87.0
ptumor	Lung	116	705	85.9
	Breast	17	95	84.8
	Gastrointestinal	9	66	88.0
	Kidney	0	32	100.0
	Other	3	30	90.9
status	Not controlled	103	634	86.0
	Controlled	42	294	87.5
neuro	No	105	656	86.2
	Yes	40	272	87.2
Total		145	928	86.5

Note: *ntumor*, number of tumors; *kps*, Karnofsky performance status; *diameter*, maximum diameter of largest tumor; *volume*, cumulative tumor volume; *ptumor*, primary tumor category; *status*, extracerebral disease status; *neuro*, neurological symptoms.

Table II. The top five models based on AIC* for the study data.

Model									AIC*
age	*sex*	*kps*	*ntumor*	*diameter*	*volume*	*ptumor*	*status*	*neuro*	1731.50
age	*sex*		*ntumor*	*diameter*	*volume*	*ptumor*	*status*	*neuro*	1731.80
age	*sex*	*kps*	*ntumor*	*diameter*		*ptumor*	*status*	*neuro*	1731.85
age	*sex*	*kps*	*ntumor*		*volume*	*ptumor*	*status*	*neuro*	1731.85
age	*sex*		*ntumor*		*volume*	*ptumor*	*status*	*neuro*	1732.16

Note: *ntumor*, number of tumors; *kps*, Karnofsky performance status; *diameter*, maximum diameter of largest tumor; *volume*, cumulative tumor volume; *ptumor*, primary tumor category; *status*, extracerebral disease status; *neuro*, neurological symptoms; AIC, Akaike's information criterion.

3. Infomation criteria for Firth's penalized partial likelihood approach

3.1. *Cox regression model*

We consider Cox regression models [1] with Andersen and Gill's [12] counting process formulation. A triplet $(\Omega, \mathcal{F}, P)$ is a probability space, and $\{\mathcal{F}_t, t \in [0,1]\}$ is an increasing right continuous family of sub σ-algebras of $\mathcal{F}$ that includes failure time and covariate histories to scaled time t and censoring histories to t^+. Let $\mathbf{N} = (N_1, \dots, N_i, \dots, N_n)^{\mathrm{T}}$ for $i = 1, \dots, n$ be an n-component multivariate counting process, where $N_i(t)$ counts the number of failures (0 or 1) for the ith individual in scaled time $t \in [0, 1]$. The sample paths $N_1, \dots, N_i, \dots, N_n$ are step functions, with 0 at time 0 and no two components having simultaneous jumps. Now, suppose that $N_i(t)$ has a random intensity process $h_i(t) = Y_i(t)h_0(t)\exp\{\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{Z}_i(t)\}$, where $h_0(t)$ is a baseline hazard function, $Y_i(t)$ is a predictable process taking the value of 1 if the ith individual is at risk at time t and 0 otherwise, $\boldsymbol{\beta}_0 = (\beta_{01}, \dots, \beta_{0p})^{\mathrm{T}}$ is a p-dimensional vector of the true regression parameters, and the p-dimensional vector $\mathbf{Z}_i = (Z_{i1}, \dots, Z_{ip})^{\mathrm{T}}$ is the predictable covariate process for the

Table III. Parameter estimates of the best model based on AIC* for the study data.

Covariate	Usual Cox regression model						Firth's penalized partial likelihood approach				
	HR	SE	95% CI		*p* (Wald)	*p* (LR)	*HR*	SE	95% CI		*p* (LR)
age (≥65 vs. <65)	1.01	0.17	0.72	1.40	0.98	0.98	1.00	0.17	0.72	1.40	0.98
sex (male vs. female)	1.20	0.19	0.83	1.73	0.34	0.33	1.19	0.19	0.83	1.72	0.34
ntumor											
1 vs. 2–4	0.72	0.20	0.49	1.06	0.09	0.09	0.72	0.20	0.49	1.06	0.09
5–10 vs. 2–4	1.58	0.22	1.03	2.41	0.04	0.04	1.59	0.22	1.04	2.43	0.04
kps (≥80 vs. <70)	1.04	0.32	0.55	1.96	0.90	0.90	1.07	0.32	0.57	2.00	0.83
diameter (≥1.6 vs. <1.6)	0.90	0.33	0.47	1.69	0.73	0.73	0.90	0.33	0.47	1.70	0.74
volume (≥1.9 vs. <1.9)	1.10	0.32	0.59	2.08	0.76	0.76	1.11	0.32	0.59	2.08	0.76
ptumor											
Breast vs. lung	1.03	0.29	0.58	1.83	0.91	0.91	1.05	0.29	0.60	1.87	0.85
GI vs. lung	1.55	0.37	0.75	3.21	0.24	0.26	1.61	0.37	0.79	3.31	0.21
Kidney vs. lung	0.00	543.30	0.00	–	0.98	<0.01	0.12	1.43	0.01	1.91	0.02
Others vs. lung	0.77	0.59	0.24	2.45	0.66	0.64	0.89	0.55	0.30	2.63	0.83
status (nc. vs. controlled)	1.02	0.19	0.71	1.47	0.92	0.92	1.03	0.19	0.71	1.48	0.89
neuro (yes vs. no)	1.50	0.23	0.96	2.36	0.07	0.08	1.51	0.23	0.97	2.37	0.07

Note: HR, hazard ratio; *kps*, Karnofsky performance status; *ntumor*, number of tumors; *diameter*, maximum diameter of largest tumor; *volume*, cumulative tumor volume; *ptumor*, primary tumor category; *status*, extracerebral disease status; *neuro*, neurological symptoms; GI, gastrointestinal; nc., not controlled; LR, likelihood ratio; SE, standard error.

ith individual. Note that the superscript 'T' indicates the transpose of a matrix or a vector. We assume that $(N_i, Y_i, \mathbf{Z}_i)$ are independent and identically distributed. In this case, the processes $M_i(t) = N_i(t) - \int_0^t h_i(x)\,\mathrm{d}x$ are independent local square integrable martingales on the scaled time interval $[0, 1]$. N_i, Y_i, and $\mathbf{Z}_i$ are assumed to be adapted to $\{\mathcal{F}_t, t \in [0, 1]\}$.

Under these settings, the log-partial likelihood function is defined as

$$l(\mathbf{N}, \mathbf{Y}, \mathbf{Z}; \boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^1 \boldsymbol{\beta}^{\mathrm{T}} \mathbf{Z}_i(x) - \log\left\{nS^{(0)}(\boldsymbol{\beta}, x)\right\} \,\mathrm{d}N_i(x),$$

the score function is defined as

$$\mathbf{U}(\boldsymbol{\beta}) = \frac{\partial l(\mathbf{N}, \mathbf{Y}, \mathbf{Z}; \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \int_0^1 \left\{ \mathbf{Z}_i(x) - \frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, x)}{S^{(0)}(\boldsymbol{\beta}, x)} \right\} \mathrm{d}N_i(x),$$

and the observed information matrix is defined as

$$\mathbf{I}(\boldsymbol{\beta}) = \frac{\partial^2 l(\mathbf{N}, \mathbf{Y}, \mathbf{Z}; \boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}} = \sum_{i=1}^{n} \int_0^1 \left[\frac{\mathbf{S}^{(2)}(\boldsymbol{\beta}, x)}{S^{(0)}(\boldsymbol{\beta}, x)} - \left\{ \frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, x)}{S^{(0)}(\boldsymbol{\beta}, x)} \right\}^{\otimes 2} \right] \mathrm{d}N_i(x),$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)^{\mathrm{T}}$, the scalar function $S^{(0)}$, the vector function $\mathbf{S}^{(1)}$, and the matrix function $\mathbf{S}^{(2)}$ are defined as $\mathbf{S}^{(k)}(\boldsymbol{\beta}, t) = n^{-1} \sum_{i=1}^{n} Y_i(t) \mathbf{Z}_i(t)^{\otimes k} \exp\{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{Z}_i(t)\}$ for $k = 0, 1, 2$. Here, for a vector $\mathbf{b}$, $\mathbf{b}^{\otimes 0} = 1$, $\mathbf{b}^{\otimes 1} = \mathbf{b}$, and $\mathbf{b}^{\otimes 2} = \mathbf{b}\mathbf{b}^{\mathrm{T}}$. By using this notation, the usual Cox regression estimator $\hat{\boldsymbol{\beta}}_{Cox}$ is obtained by solving $\mathbf{U}(\boldsymbol{\beta}) = \mathbf{0}$.

3.2. *Firth's bias correction method for Cox regression models*

Heinze and Schemper [3] directly applied Firth's bias correction method [4] to Cox regression models to overcome monotone likelihood. They proposed an estimation method based on the penalized log-partial likelihood, $l^*(\mathbf{N}, \mathbf{Y}, \mathbf{Z}; \boldsymbol{\beta}) = l(\mathbf{N}, \mathbf{Y}, \mathbf{Z}; \boldsymbol{\beta}) + 0.5 \log |\mathbf{I}(\boldsymbol{\beta})|$, and the modified score function $\mathbf{U}^*(\boldsymbol{\beta}) = \mathbf{U}(\boldsymbol{\beta}) + \mathbf{a}(\boldsymbol{\beta})$, where $|\mathbf{I}(\boldsymbol{\beta})|$ is the determinant of the observed information matrix, $\mathbf{a}(\boldsymbol{\beta}) = \{a_1(\boldsymbol{\beta}), \ldots, a_p(\boldsymbol{\beta})\}^{\mathrm{T}}$ are modification terms, and $a_j(\boldsymbol{\beta}) = \operatorname{tr}[\{\mathbf{I}(\boldsymbol{\beta})\}^{-1}\{\partial \mathbf{I}(\boldsymbol{\beta})/\partial \beta_j\}]/2$. The penalized partial likelihood estimator $\hat{\boldsymbol{\beta}}$ is obtained by solving $\mathbf{U}^*(\boldsymbol{\beta}) = \mathbf{0}$, which is different from the usual Cox regression estimator, $\hat{\boldsymbol{\beta}}_{Cox}$. They only assessed the empirical performance of these methods by using simulation studies. These simulation results confirm the satisfactory performance of the penalized likelihood ratio test and profile penalized likelihood confidence interval under monotone likelihood.

The modification term $\mathbf{a}(\boldsymbol{\beta})$ can be derived by using an asymptotic expansion of $\mathrm{E}[\mathbf{U}^*(\boldsymbol{\beta})]$. It will be convenient to employ the notation of Cox and Snell [13] and Firth [4]. Let $U_j(\boldsymbol{\beta}) = \partial l(\mathbf{N}, \mathbf{Y}, \mathbf{Z}; \boldsymbol{\beta})/\partial \beta_j$, $U_{jk}(\boldsymbol{\beta}) = \partial^2 l(\mathbf{N}, \mathbf{Y}, \mathbf{Z}; \boldsymbol{\beta})/\partial \beta_j \partial \beta_k$, $U_{jkl}(\boldsymbol{\beta}) = \partial^3 l(\mathbf{N}, \mathbf{Y}, \mathbf{Z}; \boldsymbol{\beta})/ \partial \beta_j \partial \beta_k \partial \beta_l$, the null cumulants are $\kappa_{j,k} = n^{-1}\mathrm{E}[U_j(\boldsymbol{\beta}_0)U_k(\boldsymbol{\beta}_0)]$, $\kappa_{jk} = n^{-1}\mathrm{E}[U_{jk}(\boldsymbol{\beta}_0)]$, $\kappa_{j,kl} = n^{-1}\mathrm{E}[U_j(\boldsymbol{\beta}_0)U_{kl}(\boldsymbol{\beta}_0)]$, $\kappa_{j,k,l} = n^{-1}\mathrm{E}[U_j(\boldsymbol{\beta}_0)U_k(\boldsymbol{\beta}_0)U_l(\boldsymbol{\beta}_0)]$, and $\kappa_{jkl} = n^{-1}\mathrm{E}[U_{jkl}(\boldsymbol{\beta}_0)]$. Based on the asymptotic expansion, the bias of the estimator of the mth regression parameter is given by

$$\mathrm{E}[n^{-1/2}(\hat{\beta}_m - \beta_{0m})] = n^{-1}\kappa^{l,j}\left\{-\frac{1}{2}\kappa^{k,l}(\kappa_{j,k,l} + \kappa_{j,kl}) + \alpha_j(\boldsymbol{\beta}_0)\right\} + O(n^{-3/2}), \tag{1}$$

where $\kappa^{k,l}$ denotes the inverse of the Fisher information matrix, Einstein summation convention is applied, and $\alpha_j(\boldsymbol{\beta}_0) = \mathrm{E}[a_j(\boldsymbol{\beta}_0)]$. From Eq. 1, if $\alpha_j(\boldsymbol{\beta}_0) = \kappa^{k,l}(\kappa_{j,k,l} + \kappa_{j,kl})/2$, then the first-order bias term disappears. Moreover, if $\kappa_{j,k} + \kappa_{jk} = 0$, $\kappa_{j,k,l} + \kappa_{j,kl} + \kappa_{k,jl} + \kappa_{l,jk} + \kappa_{jkl} = 0$, and $\kappa_{j,kl} = 0$, then

$$\alpha_j(\boldsymbol{\beta}) = \frac{1}{2}\kappa^{k,l}(\kappa_{j,k,l} + \kappa_{j,kl}) = \frac{1}{2}\kappa^{k,l}\kappa_{j,k,l} = -\frac{1}{2}\kappa^{k,l}\kappa_{jkl}.$$

From the aforementioned result and paying attention to the summation convention, the modification term can be written as

$$a_j(\boldsymbol{\beta}) = \frac{1}{2}\,\mathrm{tr}\,[\{\mathbf{I}(\boldsymbol{\beta})\}^{-1}\{\partial \mathbf{I}(\boldsymbol{\beta})/\partial \beta_j\}].$$

However, the relationships $\kappa_{j,k,l} + \kappa_{j,kl} + \kappa_{k,jl} + \kappa_{l,jk} + \kappa_{jkl} = 0$ and $\kappa_{j,kl} = 0$ are a nontrivial result in Cox regression models, and thus, they have never been evaluated. Nevertheless, we proved that these relationships are true in Cox regression models under independent and identically distributed (see Appendix A for more details).

3.3. Problem in heuristic information criteria

As noted earlier, although model selection is an important issue in data analysis, the model selection criteria for the penalized partial likelihood approach have never been studied. To our best knowledge, only the SAS PHREG procedure can be used to obtain an AIC-type heuristic information criterion, $\mathrm{AIC}^* = -2l^*(\mathbf{N}, \mathbf{Y}, \mathbf{Z}; \hat{\boldsymbol{\beta}}) + 2p = -2l(\mathbf{N}, \mathbf{Y}, \mathbf{Z}; \hat{\boldsymbol{\beta}}) + 2p - \log|\mathbf{I}(\hat{\boldsymbol{\beta}})|$. Moreover, other major statistical software (e.g., Stata and R) can only output the penalized log-partial likelihood. However, AIC* is not theoretically justified.

Now, we discuss a property of AIC*. After some algebra,

$$\mathrm{AIC}^* = -2l(\mathbf{N}, \mathbf{Y}, \mathbf{Z}; \hat{\boldsymbol{\beta}}) + (2 - \log n)p - \log\{n^{-p}|\mathbf{I}(\hat{\boldsymbol{\beta}})|\}.$$

The last term on the right-hand side $-\log\{n^{-p}|\mathbf{I}(\hat{\boldsymbol{\beta}})|\}$ converges to a constant because $n^{-1}\mathbf{I}(\hat{\boldsymbol{\beta}}) \xrightarrow{P} \boldsymbol{\Sigma}(\boldsymbol{\beta}_0)$ (see Appendix B for more details) and $|n^{-1}\mathbf{I}(\hat{\boldsymbol{\beta}})| = n^{-p}|\mathbf{I}(\hat{\boldsymbol{\beta}})| \xrightarrow{P} |\boldsymbol{\Sigma}(\boldsymbol{\beta}_0)|$ as $n \to \infty$, where $\boldsymbol{\Sigma}(\boldsymbol{\beta}_0) = \int_0^1 \mathbf{v}(\boldsymbol{\beta}_0, x)s^{(0)}(\boldsymbol{\beta}_0, x)h_0(x)\,\mathrm{d}x$, $\mathbf{v} = (\mathbf{s}^{(2)}/s^{(0)}) - \{\mathbf{s}^{(1)}/s^{(0)}\}^{\otimes 2}$, $s^{(0)}(\boldsymbol{\beta}, t) = \mathrm{E}[S^{(0)}(\boldsymbol{\beta}, t)]$, $\mathbf{s}^{(1)}(\boldsymbol{\beta}, t) = \mathrm{E}[\mathbf{S}^{(1)}(\boldsymbol{\beta}, t)]$, and $\mathbf{s}^{(2)}(\boldsymbol{\beta}, t) = \mathrm{E}[\mathbf{S}^{(2)}(\boldsymbol{\beta}, t)]$. If $n \geq 8$, then $2 - \log n$ is negative. Because AIC* includes the term $(2 - \log n)p$, this criterion tends to select models with large p as $n \to \infty$. Importantly, this result indicates that AIC* does not avoid over-fitting.

Similarly, the SAS PHREG procedure implements a BIC-type heuristic information criterion, $\mathrm{BIC}^* = -2l^*(\mathbf{N}, \mathbf{Y}, \mathbf{Z}; \hat{\boldsymbol{\beta}}) + p\log d$, where d is the number of events [14]. Let $c = 1 - d/n \in (0, 1]$ be the proportion of censoring, $\mathrm{BIC}^* = -2l(\mathbf{N}, \mathbf{Y}, \mathbf{Z}; \hat{\boldsymbol{\beta}}) + p\log(1 - c) - \log\{n^{-p}|\mathbf{I}(\hat{\boldsymbol{\beta}})|\}$. Because $\log(1 - c) < 0$, BIC* has a negative penalty term in proportion to the number of regression parameters.

3.4. Proposed criteria

As an alternative approach to address the issue discussed in Section 3.3, we propose a criterion termed herein AIC for Firth's penalized partial likelihood approach (AICF). AIC is a model selection criterion used to measure the goodness of fit of a model by using the risk function based on Kullback–Leibler

(KL) information between the true model and the candidate model, which is a measure of discrepancy from the true model.

Xu *et al.* [15] provided a theoretical justification for the use of partial likelihood in AIC under usual Cox regression models, which was also extended to proportional hazards mixed models. These authors developed a profile AIC [16] for selecting a model with minimum KL information based on the profile likelihood under Cox regression models. It is well known that partial likelihood can be considered as profile likelihood [17–19]. Suppose that f denotes the true distribution and $g_{\boldsymbol{\beta},\lambda} = g(\cdot;\boldsymbol{\beta},\lambda)$ denotes candidate models, where $\lambda \in \boldsymbol{\Lambda}$ is the nuisance parameter and $\boldsymbol{\Lambda}$ is the parameter space of λ. The KL information can be written as $KL(f, g_{\boldsymbol{\beta},\lambda}) = \mathrm{E}_{(N,Y,\mathbf{Z})\sim f}[\log f(N,Y,\mathbf{Z}) - \log g_{\boldsymbol{\beta},\lambda}(N,Y,\mathbf{Z})]$. Here and subsequently, we write $\mathrm{E}_N = \mathrm{E}_{(N,Y,\mathbf{Z})\sim f}$ for convenience. Focusing on the regression parameters $\boldsymbol{\beta}$ alone and ignoring the constant term $\mathrm{E}_N[\log f(N,Y,\mathbf{Z})]$ in KL, the minimum KL information is given at $\boldsymbol{\beta}_0$ such that $\mathrm{E}_N[\log g_{\boldsymbol{\beta}_0}(N,Y,\mathbf{Z})] = \max_{\boldsymbol{\beta}} \mathrm{E}_N[\log g_{\boldsymbol{\beta}}(N,Y,\mathbf{Z})]$, where $\mathrm{E}_N[\log g_{\boldsymbol{\beta}}(N,Y,\mathbf{Z})] = \max_\lambda \mathrm{E}_N[\log g_{\boldsymbol{\beta},\lambda}(N,Y,\mathbf{Z})]$. If the model is correctly specified (i.e., $f = g_{\boldsymbol{\beta}_0}$), $\mathrm{E}_N[\log g_{\boldsymbol{\beta}_0}(N,Y,\mathbf{Z})] = \int_\Omega \log g_{\boldsymbol{\beta}_0}(N,Y,\mathbf{Z})\,\mathrm{d}P$. Under Cox regression models, the log profile likelihood can be written as $\max_\lambda \sum_{i=1}^n g_{\boldsymbol{\beta}}(N_i, Y_i, \mathbf{Z}_i;\boldsymbol{\beta},\lambda) = l(\mathbf{N},\mathbf{Y},\mathbf{Z};\boldsymbol{\beta})$. Xu *et al.* [15] showed that the risk function, $\mathrm{E}_N\mathrm{E}_{\tilde{N}}[-2l(\tilde{\mathbf{N}},\tilde{\mathbf{Y}},\tilde{\mathbf{Z}};\hat{\boldsymbol{\beta}}_{Cox})]$ based on the log profile likelihood, and profile AIC, $-2l(\tilde{\mathbf{N}},\tilde{\mathbf{Y}},\tilde{\mathbf{Z}};\hat{\boldsymbol{\beta}}_{Cox}) + 2p$, as an approximately unbiased estimator of the risk function, where $(\tilde{\mathbf{N}},\tilde{\mathbf{Y}},\tilde{\mathbf{Z}})$ is a future observation. Based on Akaike [8], the minimum risk function corresponds to the minimum KL information using a future observation.

Therefore, we consider a partial likelihood-based risk function, $RISK = \mathrm{E}_N\mathrm{E}_{\tilde{N}}[-2l(\tilde{\mathbf{N}},\tilde{\mathbf{Y}},\tilde{\mathbf{Z}};\hat{\boldsymbol{\beta}})]$, and derive AICF as an approximately unbiased estimator of $RISK$. In the definition of $RISK$, the estimator $\hat{\boldsymbol{\beta}}$ was not the usual Cox regression estimator $\hat{\boldsymbol{\beta}}_{Cox}$, but rather the Firth's penalized partial likelihood estimator. When we simply estimate $RISK$ by $-2l(\mathbf{N},\mathbf{Y},\mathbf{Z};\hat{\boldsymbol{\beta}})$, we need to correct bias B. Here, B is defined as

$$\begin{aligned} B &= \mathrm{E}_N\mathrm{E}_{\tilde{N}}[2l(\mathbf{N},\mathbf{Y},\mathbf{Z};\hat{\boldsymbol{\beta}}) - 2l(\tilde{\mathbf{N}},\tilde{\mathbf{Y}},\tilde{\mathbf{Z}};\hat{\boldsymbol{\beta}})] \\ &= b_1 + b_2 + b_3, \end{aligned}$$

where

$$\begin{aligned} b_1 &= \mathrm{E}_N[\mathrm{E}_{\tilde{N}}[2l(\tilde{\mathbf{N}},\tilde{\mathbf{Y}},\tilde{\mathbf{Z}};\boldsymbol{\beta}_0)] - \mathrm{E}_{\tilde{N}}[2l(\tilde{\mathbf{N}},\tilde{\mathbf{Y}},\tilde{\mathbf{Z}};\hat{\boldsymbol{\beta}})]], \\ b_2 &= \mathrm{E}_N[2l(\mathbf{N},\mathbf{Y},\mathbf{Z};\boldsymbol{\beta}_0)] - \mathrm{E}_{\tilde{N}}[2l(\tilde{\mathbf{N}},\tilde{\mathbf{Y}},\tilde{\mathbf{Z}};\boldsymbol{\beta}_0)], \end{aligned}$$

and

$$b_3 = \mathrm{E}_N[2l(\mathbf{N},\mathbf{Y},\mathbf{Z};\hat{\boldsymbol{\beta}}) - 2l(\mathbf{N},\mathbf{Y},\mathbf{Z};\boldsymbol{\beta}_0)].$$

According to this definition, B includes the true parameter vector $\boldsymbol{\beta}_0$. Therefore, we need approximate B. A second-order Taylor expansion of $\mathrm{E}_{\tilde{N}}[l(\tilde{\mathbf{N}},\tilde{\mathbf{Y}},\tilde{\mathbf{Z}};\hat{\boldsymbol{\beta}})]$ around $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0$ gives

$$\mathrm{E}_{\tilde{N}}[l(\tilde{\mathbf{N}},\tilde{\mathbf{Y}},\tilde{\mathbf{Z}};\hat{\boldsymbol{\beta}})] \approx \mathrm{E}_{\tilde{N}}[l(\tilde{\mathbf{N}},\tilde{\mathbf{Y}},\tilde{\mathbf{Z}};\boldsymbol{\beta}_0)] - \frac{1}{2}\{\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\mathrm{T}}\}\boldsymbol{\Sigma}(\boldsymbol{\beta}_0)\{\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\}, \tag{2}$$

a first-order Taylor expansion of $\mathbf{U}^*(\hat{\boldsymbol{\beta}}) = \mathbf{0}$ around $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0$ gives

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \approx \{n^{-1}\mathbf{I}^*(\boldsymbol{\beta}_0)\}^{-1}\{n^{-1/2}\mathbf{U}^*(\boldsymbol{\beta}_0)\}, \tag{3}$$

and a second-order Taylor expansion of $l(\mathbf{N},\mathbf{Y},\mathbf{Z};\boldsymbol{\beta}_0)$ around $\boldsymbol{\beta}_0 = \hat{\boldsymbol{\beta}}$ gives

$$l(\mathbf{N},\mathbf{Y},\mathbf{Z};\boldsymbol{\beta}_0) \approx l(\mathbf{N},\mathbf{Y},\mathbf{Z};\hat{\boldsymbol{\beta}}) - \frac{1}{2}\{\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\mathrm{T}}\}\{n^{-1}\mathbf{I}(\hat{\boldsymbol{\beta}})\}\{\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\} + \{\mathbf{a}(\hat{\boldsymbol{\beta}})\}^{\mathrm{T}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0), \tag{4}$$

where $\mathbf{I}^*(\boldsymbol{\beta}) = -\partial\mathbf{U}^*(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^{\mathrm{T}} = \mathbf{I}(\boldsymbol{\beta}) - \partial\mathbf{a}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^{\mathrm{T}}$. From the fact $\mathrm{E}[f(\mathbf{D})] = \mathrm{E}[\ \mathrm{tr}\ \{f(\mathbf{D})\}]$ for a scalar function f and a random vector $\mathbf{D}$, and by substituting Eqs 2-4 into b_1, we can show that

$$b_1 \approx \mathrm{E}_N[\ \mathrm{tr}\ \{\boldsymbol{\Sigma}(\boldsymbol{\beta}_0)\{n^{-1}\mathbf{I}^*(\boldsymbol{\beta}_0)\}^{-1}\{n^{-1}\mathbf{J}^*(\boldsymbol{\beta}_0)\}\{n^{-1}\mathbf{I}^*(\boldsymbol{\beta}_0)\}^{-1}\}],$$

where $\mathbf{J}^*(\boldsymbol{\beta}) = \sum_{i=1}^n\{\mathbf{L}_i^*(\boldsymbol{\beta})\}^{\otimes 2}$, $\mathbf{L}_i^*(\boldsymbol{\beta}) = \mathbf{L}_i(\boldsymbol{\beta}) - \mathbf{a}(\boldsymbol{\beta})/n$, and $\mathbf{L}_i(\boldsymbol{\beta}_0) = \int_0^1\{\mathbf{Z}_i(x) - \mathbf{S}^{(1)}(\boldsymbol{\beta}_0,x)/S^{(0)}(\boldsymbol{\beta}_0,x)\}\,\mathrm{d}M_i(x)$. Similarly,

$$b_3 \approx \mathrm{E}_N[\ \mathrm{tr}\ \{\{n^{-1}\mathbf{I}(\hat{\boldsymbol{\beta}})\}\{n^{-1}\mathbf{I}^*(\boldsymbol{\beta}_0)\}^{-1}\{n^{-1}\mathbf{J}^*(\boldsymbol{\beta}_0)\}\{n^{-1}\mathbf{I}^*(\boldsymbol{\beta}_0)\}^{-1} - 2\{\mathbf{a}(\hat{\boldsymbol{\beta}})\}^{\mathrm{T}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\}].$$

Under the true model $n^{-1}\mathbf{J}^*(\boldsymbol{\beta}_0) \xrightarrow{P} \boldsymbol{\Sigma}(\boldsymbol{\beta}_0)$, $n^{-1}\mathbf{I}^*(\boldsymbol{\beta}_0) \xrightarrow{P} \boldsymbol{\Sigma}(\boldsymbol{\beta}_0)$, $n^{-1}\mathbf{I}(\hat{\boldsymbol{\beta}}) \xrightarrow{P} \boldsymbol{\Sigma}(\boldsymbol{\beta}_0)$, and $\mathbf{a}(\hat{\boldsymbol{\beta}}) = O_p(1)$ (see Data S1 and Appendix B). Therefore, by applying the continuous mapping theorem, we obtain

$$\begin{aligned} &\mathrm{tr}\,\{\boldsymbol{\Sigma}(\boldsymbol{\beta}_0)\{n^{-1}\mathbf{I}^*(\boldsymbol{\beta}_0)\}^{-1}\{n^{-1}\mathbf{J}^*(\boldsymbol{\beta}_0)\}\{n^{-1}\mathbf{I}^*(\boldsymbol{\beta}_0)\}^{-1}\}\\ &\quad\xrightarrow{P} \mathrm{tr}\,\{\boldsymbol{\Sigma}(\boldsymbol{\beta}_0)\{\boldsymbol{\Sigma}(\boldsymbol{\beta}_0)\}^{-1}\boldsymbol{\Sigma}(\boldsymbol{\beta}_0)\{\boldsymbol{\Sigma}(\boldsymbol{\beta}_0)\}^{-1}\} = p, \end{aligned}$$

and

$$\begin{aligned} &\mathrm{tr}\,\{\{n^{-1}\mathbf{I}(\hat{\boldsymbol{\beta}})\}\{n^{-1}\mathbf{I}^*(\boldsymbol{\beta}_0)\}^{-1}\{n^{-1}\mathbf{J}^*(\boldsymbol{\beta}_0)\}\{n^{-1}\mathbf{I}^*(\boldsymbol{\beta}_0)\}^{-1} - 2\{\mathbf{a}(\hat{\boldsymbol{\beta}})\}^{\mathrm{T}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\}\\ &\quad\xrightarrow{P} \mathrm{tr}\,\{\boldsymbol{\Sigma}(\boldsymbol{\beta}_0)\{\boldsymbol{\Sigma}(\boldsymbol{\beta}_0)\}^{-1}\boldsymbol{\Sigma}(\boldsymbol{\beta}_0)\{\boldsymbol{\Sigma}(\boldsymbol{\beta}_0)\}^{-1}\} = p, \end{aligned}$$

as $n \to \infty$. Moreover, it is obvious that $b_2 = 0$. Further details are presented in Data S1. Hence, $b_1 \approx p$, $b_3 \approx p$, and bias $B = b_1 + b_2 + b_3$ can be approximated by $2p$. From the aforementioned results, we define AICF as

$$\text{AICF} = -2l(\mathbf{N}, \mathbf{Y}, \mathbf{Z}; \hat{\boldsymbol{\beta}}) + 2p.$$

The AICF does not include the penalty term of AIC*, $0.5 \log |\mathbf{I}(\hat{\boldsymbol{\beta}})|$. Even in the penalized partial likelihood setting, non-penalized likelihood should be used for risk estimation. Sometimes, penalty terms for parameter estimation have a strong effect on a model selection criterion.

Similarly, based on the results of a previous study [14], we propose BIC for Firth's penalized partial likelihood approach

$$\text{BICF} = -2l(\mathbf{N}, \mathbf{Y}, \mathbf{Z}; \hat{\boldsymbol{\beta}}) + p \log d.$$

The detailed derivation is omitted, but is similar to that described previously [14].

Note that SAS and R programs for AICF and BICF are provided in Data S3.

4. Simulation

4.1. Simulation conditions

Simulation studies were conducted to investigate the performance of model selection critera (AICF, BICF, AIC*, and BIC*) in a monotone likelihood setting and to check that the property of AIC* discussed in Section 3.3 holds. We set the simulation conditions by referring to [3] and the generated observations $\{N_i, Y_i, \mathbf{Z}_i\}$ from exponential distributions with hazard functions $h_i(t) = h_0(t)\gamma_i(t) = h_0(t)\exp\{\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{Z}_i\}$, where $h_0(t) = 1$, $\boldsymbol{\beta}_0 = (\log\theta, \log\theta, \log\theta, 0, 0)^{\mathrm{T}}$, $\mathbf{Z}_i = (Z_{i1}, Z_{i2}, Z_{i3}, Z_{i4}, Z_{i5})^{\mathrm{T}}$, and $Z_{ij} \sim \text{Bernoulli}\ (q)$. We further set the proportion of covariates as $q = 0.5$ or 0.8, the regression parameters as $\theta = 1.3, 2, 4$, or 16, the proportion of censoring as $c = 0, 50$, or 90 (%), and the total sample size as $n = 100, 200$, or 1000. We generated data under simple type I censoring; the observations of each individual were censored at a suitable time τ for each simulation. Time τ was determined to achieve an expected 50% and 90% censoring. We find monotone likelihood in situations of high censoring and high parameter values. For each data configuration, we generated $R = 20,000$ simulations. For each simulation, we calculated AICF, AIC*, BICF, and BIC* for the following 11 models:

Model 1: $\log\gamma_i(t) = \beta_1 Z_{i1}$
Model 2: $\log\gamma_i(t) = \beta_4 Z_{i4}$
Model 3: $\log\gamma_i(t) = \beta_1 Z_{i1} + \beta_2 Z_{i2}$
Model 4: $\log\gamma_i(t) = \beta_1 Z_{i1} + \beta_4 Z_{i4}$
Model 5: $\log\gamma_i(t) = \beta_4 Z_{i4} + \beta_5 Z_{i5}$
Model 6: $\log\gamma_i(t) = \beta_1 Z_{i1} + \beta_2 Z_{i2} + \beta_3 Z_{i3}$ **(the true model)**
Model 7: $\log\gamma_i(t) = \beta_1 Z_{i1} + \beta_2 Z_{i2} + \beta_4 Z_{i4}$
Model 8: $\log\gamma_i(t) = \beta_1 Z_{i1} + \beta_4 Z_{i4} + \beta_5 Z_{i5}$
Model 9: $\log\gamma_i(t) = \beta_1 Z_{i1} + \beta_2 Z_{i2} + \beta_3 Z_{i3} + \beta_4 Z_{i4}$
Model 10: $\log\gamma_i(t) = \beta_1 Z_{i1} + \beta_2 Z_{i2} + \beta_4 Z_{i4} + \beta_5 Z_{i5}$
Model 11: $\log\gamma_i(t) = \beta_1 Z_{i1} + \beta_2 Z_{i2} + \beta_3 Z_{i3} + \beta_4 Z_{i4} + \beta_5 Z_{i5}$ **(the full model)**

Model 6 is the true model, and Model 11 is the full model, that is, the model with maximum p.

Because it is well known that AIC is designed for optimal prediction and BIC is designed to identify the true model, we assessed the predictive performance of AICF and AIC*. We evaluated the mean of the difference between the estimated mean risk (MR) and value of the information criterion in each model and its 5 and 95 percentiles, as well as the estimated MR for the selected model based on new data. The MR and its estimator, $\widehat{\mathrm{MR}}$, are defined as

$$\mathrm{MR} = \frac{1}{R}\sum_{r=1}^{R} \mathrm{E}_{\tilde{N}}\left[-2l(\tilde{\mathbf{N}}, \tilde{\mathbf{Y}}, \tilde{\mathbf{Z}}; \hat{\boldsymbol{\beta}}_r)\right]$$

and

$$\widehat{\mathrm{MR}} = -\frac{2}{R}\sum_{r=1}^{R}\sum_{i=1}^{n}\int_0^1 \left[\hat{\boldsymbol{\beta}}_r^{\mathrm{T}}\tilde{\mathbf{Z}}_{ri}(x) - \log\left\{\sum_{j=1}^{n}\tilde{Y}_{rj}(x)\exp\{\hat{\boldsymbol{\beta}}_r^{\mathrm{T}}\tilde{\mathbf{Z}}_{rj}(x)\}\right\}\right]\mathrm{d}\tilde{N}_{ri}(x),$$

where $(\tilde{\mathbf{N}}, \tilde{\mathbf{Y}}, \tilde{\mathbf{Z}})$ is another dataset of the same size, $\hat{\boldsymbol{\beta}}_r$ is the model estimate of each replication, and $\mathbf{Z}_{ri}$ is the covariate vector in each model and replication. The absolute value of the mean difference between the $\widehat{\mathrm{MR}}$ and the value of the information criterion should be small because AIC is an estimator of risk function; thus, the mean difference can be regarded as an empirical bias of an AIC-type criterion. The estimated MR for the selected model, $\widehat{\mathrm{MR}}_{sel}$, is defined as

$$\widehat{\mathrm{MR}}_{sel} = -\frac{2}{R}\sum_{r=1}^{R}\sum_{i=1}^{n}\int_0^1 \left[\hat{\boldsymbol{\beta}}_{r,sel}^{\mathrm{T}}\tilde{\mathbf{Z}}_{ri}(x) - \log\left\{\sum_{j=1}^{n}\tilde{Y}_{rj}(x)\exp\{\hat{\boldsymbol{\beta}}_{r,sel}^{\mathrm{T}}\tilde{\mathbf{Z}}_{rj}(x)\}\right\}\right]\mathrm{d}\tilde{N}_{ri}(x),$$

where $\hat{\boldsymbol{\beta}}_{r,sel}$ is an estimate of the selected model. The $\widehat{\mathrm{MR}}_{sel}$ should be small and is considered to be a performance indicator for prediction, as it measures the goodness of fit for the selected model to a future data as a mean deviance.

To assess the performance of the criteria, we evaluated model selection probability P_m, where $m = 1, 2, \dots, 11$. We estimate the model selection probability by $\hat{P}_m = \{\#\text{ of the model } m \text{ selected}\}/\{\#\text{ of replication } (R = 20,000)\}$, the relative frequency of the model obtained by minimizing the information criterion.

4.2. Simulation results

The mean of the difference between the estimated MR and value of information criterion and its 5 and 95 percentiles for Models 6 and 11 for $q = 0.5$ is shown in Table IV. The mean differences for AICF were smaller than those for AIC* under all conditions. The mean differences for AIC* increased with the number of events. The 5 and 95 percentiles for AICF were approximately symmetric around 0, whereas the percentiles for AIC* were not symmetric. For instance, in the case with $c = 0\%$, $\theta = 1$, $n = 100$, and Model 6 (true model), the mean difference and its percentiles for AICF and AIC* were −0.4 (−9.9, 5.9) and −9.8 (−19.2, −3.5); in the case with $c = 0\%$, $\theta = 1$, $n = 100$, and Model 11 (full model), the mean difference and its percentiles for AICF and AIC* were −1.1 (−11.2, 6.0) and −16.7 (−26.7, −9.7). Thus, AIC* is clearly biased downward, as AIC* includes unnecessary negative terms (Section 3.3). Larger bias was observed in models with large p-values. Therefore, these models cannot estimate the risk function. The estimated MR for the selected model based on another new dataset, $\widehat{\mathrm{MR}}_{sel}$, for $q = 0.5$ is shown in Table IV. The $\widehat{\mathrm{MR}}_{sel}$ for AICF was smaller than that for AIC*, except in a case with $c = 50\%$, $\theta = 4$, and $n = 1000$, and $c = 0\%$, $\theta = 1$, and $n = 200$. Thus, the model selection based on AICF showed small deviance for future data. These results revealed that AICF shows better prediction performance.

The selection probability of Models 6 (the true model) and 11 (the full model) for AICF and AIC* when $q = 0.5$ is shown in Table V. For larger parameter values, larger sample sizes, and less censoring, the selection probability of the true model is larger for AICF than for AIC*. For smaller parameter values, smaller sample sizes, and more censoring, the selection probability of the true model is larger for AIC* than for AICF. The selection probability of the full model for AICF is smaller than that for AIC*. On the contrary, the selection probability of the full model for AIC* increases with the number of events. In particular, for $n = 1000$, the selection probability of the full model for AIC* is equal to one because of the term $(2 - \log n)p$ in AIC*, as discussed in Section 3.3.

Table IV. The mean of the difference between the estimated MR and the value of the information criterion in each model and its 5 and 95 percentiles, and the estimated MR for the selected model (the proportion of covariates: $q = 0.5$; the number of simulations: $R = 20,000$).

c (%)	θ	n	Mean difference (5 percentile, 95 percentile)				$\widehat{\mathrm{MR}}_{sel}$	
			Model 6 (true model)		Model 11 (full model)			
			AICF	AIC*	AICF	AIC*	AICF	AIC*
90	1	100	**−0.4 (−43.9, 45.9)**	−2.7 (−44.2, 42.1)	**−1.1 (−44.4, 45.5)**	−4.8 (−44.7, 39.2)	**94.38**	95.05
90	1	200	**0.1 (−71.4, 74.6)**	−4.4 (−74.8, 68.9)	**−0.1 (−71.6, 74.9)**	−7.7 (−77.1, 65.5)	**212.87**	214.55
90	1	1000	**1.6 (−206.1, 218.1)**	−8.0 (−215.2, 208.0)	**1.6 (−206.9, 218.2)**	−14.4 (−222.1, 201.4)	**1373.64**	1375.61
90	2	100	**−0.4 (−43.8, 45.8)**	−2.7 (−44.1, 42.0)	**−1.0 (−44.2, 45.6)**	−4.7 (−44.6, 39.3)	**94.40**	95.04
90	2	200	**−0.6 (−72.0, 74.4)**	−5.2 (−75.4, 68.8)	**−0.9 (−72.4, 74.6)**	−8.5 (−77.8, 65.2)	**212.82**	214.48
90	2	1000	**1.3 (−205.9, 217.1)**	−8.3 (−215.1, 207.0)	**1.3 (−206.6, 217.4)**	−14.8 (−221.8, 200.6)	**1375.21**	1377.21
90	4	100	**−0.7 (−44.2, 45.7)**	−3.0 (−44.5, 41.9)	**−1.4 (−44.6, 45.4)**	−5.1 (−45.0, 39.1)	**94.73**	95.36
90	4	200	**0.0 (−71.6, 74.2)**	−4.6 (−74.9, 68.6)	**−0.3 (−71.5, 74.3)**	−7.9 (−77.1, 65.0)	**212.88**	214.50
90	4	1000	**−1.8 (−216.9, 214.8)**	−11.4 (−226.1, 204.7)	**−1.9 (−215.8, 213.4)**	−17.9 (−231.0, 196.7)	**1376.07**	1378.04
90	16	100	**−0.7 (−43.4, 45.9)**	−3.0 (−43.8, 42.1)	**−1.4 (−44.0, 45.2)**	−5.1 (−44.4, 38.9)	**94.49**	95.15
90	16	200	**−0.6 (−71.3, 74.0)**	−5.2 (−74.7, 68.3)	**−0.9 (−71.5, 74.0)**	−8.5 (−77.1, 64.6)	**212.88**	214.54
90	16	1000	**0.1 (−207.5, 217.3)**	−9.5 (−216.6, 207.3)	**0.1 (−207.6, 217.4)**	−15.9 (−222.8, 200.6)	**1374.48**	1376.42
50	1	100	**0.1 (−64.8, 62.8)**	−7.4 (−71.8, 54.9)	**−0.4 (−65.8, 62.3)**	−12.8 (−77.4, 49.1)	**432.37**	434.56
50	1	200	**−1.1 (−110.0, 104.1)**	−10.7 (−119.2, 94.2)	**−1.3 (−109.7, 104.7)**	−17.2 (−125.1, 88.2)	**1000.80**	1002.75
50	1	1000	**3.0 (−322.5, 325.7)**	−11.5 (−336.8, 311.1)	**2.9 (−322.5, 325.9)**	−21.2 (−346.3, 301.5)	**6602.19**	6603.65
50	2	100	**−0.4 (−64.5, 63.2)**	−7.8 (−71.4, 55.2)	**−0.9 (−65.6, 62.8)**	−13.3 (−77.1, 49.7)	**432.28**	434.21
50	2	200	**0.9 (−108.1, 107.4)**	−8.7 (−117.3, 97.4)	**0.6 (−108.1, 106.9)**	−15.3 (−123.5, 90.4)	**999.82**	1001.35
50	2	1000	**−2.9 (−322.7, 319.4)**	−17.4 (−337.0, 304.8)	**−3.0 (−322.5, 318.6)**	−27.1 (−346.4, 294.3)	**6601.73**	6601.85
50	4	100	**−0.9 (−66.0, 61.9)**	−8.4 (−73.0, 54.0)	**−1.6 (−66.9, 61.3)**	−14.0 (−78.4, 48.2)	**432.62**	434.21
50	4	200	**0.1 (−107.2, 104.3)**	−9.5 (−116.4, 94.4)	**−0.3 (−107.3, 103.9)**	−16.3 (−122.6, 87.4)	**999.69**	1000.66
50	4	1000	**−3.0 (−323.1, 312.3)**	−17.4 (−337.4, 297.7)	**−3.2 (−323.1, 311.9)**	−27.3 (−347.0, 287.6)	6595.55	**6595.42**
50	16	100	**0.1 (−64.3, 63.4)**	−7.3 (−71.2, 55.5)	**−0.7 (−65.6, 62.4)**	−13.1 (−77.0, 49.3)	**431.82**	433.14
50	16	200	**−1.1 (−107.8, 104.4)**	−10.6 (−117.0, 94.5)	**−1.6 (−108.1, 104.4)**	−17.6 (−123.4, 87.9)	**999.22**	999.75
50	16	1000	**0.8 (−320.5, 319.0)**	−13.7 (−334.8, 304.4)	**0.5 (−321.0, 318.5)**	−23.6 (−344.8, 294.1)	**6589.13**	6589.34
0	1	100	**−0.4 (−9.9, 5.9)**	−9.8 (−19.2, −3.5)	**−1.1 (−11.2, 6.0)**	−16.7 (−26.7, −9.7)	**728.44**	728.71
0	1	200	**−0.1 (−12.1, 8.8)**	−11.7 (−23.7, −2.8)	**−0.5 (−13.2, 9.0)**	−19.8 (−32.4, −10.3)	1722.43	**1722.27**
0	1	1000	**0.2 (−23.8, 21.1)**	−16.2 (−40.3, 4.6)	**0.1 (−24.3, 21.3)**	−27.4 (−51.8, −6.3)	**11779.42**	11780.44
0	2	100	**−0.5 (−17.2, 14.1)**	−9.6 (−26.2, 4.7)	**−1.2 (−18.4, 13.8)**	−16.6 (−33.6, −1.7)	**705.35**	706.37
0	2	200	**−0.2 (−23.6, 20.8)**	−11.6 (−34.8, 9.3)	**−0.5 (−23.9, 20.5)**	−19.5 (−42.9, 1.4)	**1675.67**	1676.76
0	2	1000	**0.4 (−48.9, 48.0)**	−15.9 (−65.2, 31.7)	**0.3 (−49.3, 48.2)**	−27.0 (−76.6, 20.8)	**11548.20**	11549.24
0	4	100	**−0.5 (−24.1, 21.7)**	−9.2 (−32.6, 12.8)	**−1.1 (−24.9, 21.4)**	−16.0 (−39.5, 6.2)	**657.00**	658.21
0	4	200	**−0.5 (−33.3, 30.8)**	−11.4 (−44.1, 19.8)	**−0.9 (−34.0, 30.7)**	−19.5 (−52.4, 12.0)	**1578.08**	1579.22
0	4	1000	**−0.4 (−72.1, 69.9)**	−16.2 (−87.8, 54.0)	**−0.4 (−72.1, 69.9)**	−27.3 (−98.8, 43.0)	**11053.44**	11054.48
0	16	100	**−0.3 (−26.8, 26.5)**	−7.9 (−34.0, 18.4)	**−0.8 (−27.7, 26.2)**	−14.6 (−41.0, 12.0)	**575.12**	576.30
0	16	200	**−0.1 (−37.4, 37.8)**	−10.0 (−47.0, 27.6)	**−0.4 (−37.8, 37.6)**	−17.9 (−55.1, 19.8)	**1411.05**	1412.15
0	16	1000	**0.4 (−81.5, 84.4)**	−14.4 (−96.3, 69.4)	**0.4 (−82.0, 84.4)**	−25.5 (−107.8, 58.5)	**10203.57**	10204.62

MR, mean risk; Mean difference (5 percentile, 95 percentile), mean of the difference between the estimated mean risk, $\widehat{\mathrm{MR}}$, and the value of the information criterion in each model and its 5 and 95 percentiles; $\widehat{\mathrm{MR}}_{sel}$, estimated MR for the selected model based on new data; c, proportion of random censoring; θ, regression parameters; n, total sample size. The values that are superior to other are highlighted.

Table V. The selection probability (the proportion of covariates: $q = 0.5$; the number of simulations: $R = 20,000$).

c (%)	θ	n	Model 6 (true model)				Model 11 (full model)			
			AICF	AIC*	BICF	BIC*	AICF	AIC*	BICF	BIC*
90	1.3	100	0.060	**0.106**	0.050	0.087	0.008	0.039	**0.007**	0.016
90	1.3	200	0.058	**0.121**	0.026	0.094	0.005	0.239	**0.001**	0.016
90	1.3	1000	0.059	0.000	0.005	**0.094**	0.005	1.000	**0.000**	0.019
90	2	100	0.063	**0.110**	0.052	0.092	0.008	0.041	**0.007**	0.019
90	2	200	0.060	**0.116**	0.024	0.093	0.005	0.238	**0.001**	0.018
90	2	1000	0.062	0.000	0.006	**0.096**	0.004	1.000	**0.000**	0.018
90	4	100	0.061	**0.106**	0.052	0.086	0.008	0.042	**0.007**	0.018
90	4	200	0.061	**0.115**	0.026	0.092	0.006	0.239	**0.001**	0.018
90	4	1000	0.055	0.000	0.005	**0.090**	0.005	1.000	**0.000**	0.019
90	16	100	0.062	**0.110**	0.052	0.091	0.008	0.043	**0.006**	0.018
90	16	200	0.058	**0.119**	0.023	0.093	0.006	0.234	**0.001**	0.017
90	16	1000	0.060	0.000	0.005	**0.098**	0.005	1.000	**0.000**	0.020
50	1.3	100	0.062	0.000	0.011	**0.095**	0.007	1.000	**0.000**	0.021
50	1.3	200	0.064	0.000	0.006	**0.104**	0.006	1.000	**0.000**	0.022
50	1.3	1000	0.091	0.000	0.003	**0.135**	0.007	1.000	**0.000**	0.027
50	2	100	0.082	0.000	0.018	**0.122**	0.010	1.000	**0.000**	0.027
50	2	200	0.099	0.000	0.015	**0.140**	0.010	1.000	**0.000**	0.031
50	2	1000	0.258	0.000	0.028	**0.294**	0.025	1.000	**0.000**	0.070
50	4	100	0.109	0.000	0.030	**0.148**	0.017	1.000	**0.000**	0.039
50	4	200	0.154	0.000	0.031	**0.194**	0.020	1.000	**0.000**	0.052
50	4	1000	**0.469**	0.000	0.142	0.462	0.053	1.000	**0.000**	0.116
50	16	100	0.133	0.000	0.044	**0.172**	0.022	1.000	**0.002**	0.050
50	16	200	0.203	0.000	0.054	**0.242**	0.029	1.000	**0.001**	0.066
50	16	1000	**0.596**	0.000	0.305	0.544	0.072	1.000	**0.000**	0.143
0	1.3	100	0.270	0.000	0.087	**0.304**	0.033	1.000	**0.001**	0.067
0	1.3	200	**0.468**	0.000	0.188	0.468	0.054	1.000	**0.001**	0.108
0	1.3	1000	0.778	0.000	**0.919**	0.658	0.085	1.000	**0.001**	0.155
0	2	100	0.745	0.000	**0.820**	0.662	0.091	1.000	**0.007**	0.145
0	2	200	0.778	0.000	**0.968**	0.670	0.086	1.000	**0.003**	0.150
0	2	1000	0.786	0.000	**0.991**	0.670	0.081	1.000	**0.001**	0.149
0	4	100	0.772	0.000	**0.959**	0.676	0.089	1.000	**0.007**	0.144
0	4	200	0.773	0.000	**0.971**	0.664	0.085	1.000	**0.004**	0.150
0	4	1000	0.786	0.000	**0.990**	0.670	0.081	1.000	**0.001**	0.149
0	16	100	0.774	0.000	**0.958**	0.690	0.088	1.000	**0.007**	0.137
0	16	200	0.779	0.000	**0.974**	0.679	0.085	1.000	**0.003**	0.144
0	16	1000	0.788	0.000	**0.991**	0.668	0.081	1.000	**0.001**	0.151

Note: c, proportion of random censoring; θ, regression parameters; n, total sample size. The values that are superior to other are highlighted.

The selection probability of Models 6 (true model) and 11 (full model) for BICF and BIC* when $q = 0.5$ is shown in Table V. For larger parameter values, larger sample sizes, and less censoring, the selection probability of the true model was larger for BICF than for BIC*. For smaller parameter values, smaller sample sizes, and more censoring, the selection probability of the true model was larger for BIC* than for BICF. The selection probability of the full model for BICF showed the smallest value. Although a BIC-type criterion was designed to identify the true model as $n \to \infty$, the selection probability of the true model of BIC* was smaller than that of BICF for a large number of samples. For instance, in the case with $c = 0\%$, $\theta = 16$, and $n = 1000$, the selection probability of the true model for BICF and BIC* were 0.991 and 0.668, while the selection probability of the full model for BICF and BIC* were 0.001 and 0.151. This is because of the properties of BIC* discussed in Section 3.3.

The results of the other models and conditions are presented in Data S2 (Tables S1–S8 for bias, Table S9 for prediction performance, and Tables S10–S17 for selection probability); these results reveal the same tendencies as discussed previously.

In summary, AICF can estimate the risk function and shows better prediction performance than AIC*. By contrast, because AIC* selects models that have many parameters and ignores the adequacy of the model as the number of events increase, it is not as efficient and not recommended for model selection. BICF can select the true model as $n \rightarrow \infty$. In contrast, because the selection probability of the true model for BIC* was smaller than that for BICF for a large number of samples, this method is not as efficient in selecting the true model.

5. Application of the prospective observational study data

Next, we apply the proposed method to the study introduced in Section 2. We select a model based on AICF as in that section, with the result shown in Table VI. As AIC* tends to select models that have many parameters (Table II), the model that includes *ntumor* (number of tumors), *ptumor* (primary tumor category), and *neuro* (neurological symptoms) was selected by AICF as the best. The best model under AICF does not include unnecessary covariates such as *age* and *status* whose hazard ratios were almost equal to one (Table II).

The estimated hazard ratios for the best model are shown in Table VII. The best model under AICF included *ntumor* (1 vs. 2–4: $HR = 0.71$, p-value $= 0.08$; 5–10 vs. 2–4: $HR = 1.57$, p-value $= 0.04$), *ptumor* (breast vs. lung: $HR = 0.94$, p-value $= 0.82$; gastrointestinal vs. lung: $HR = 1.60$, p-value $= 0.21$; kidney vs. lung: $HR = 0.12$, p-value $= 0.02$; others vs. lung: $HR = 0.87$, p-value $= 0.79$), and *neuro* ($HR = 1.53$, p-value $= 0.04$). The selected covariates under AICF are better than *age* (Table III; $HR = 1.00$, p-value $= 0.98$) and *status* (Table III; $HR = 1.03$, p-value $= 0.85$). Therefore, the best model under AICF seems to be plausible and may be more appropriate than the best model under AIC*. Additionally, the best model under AICF is clinically interpretable because the selected variables are prognostic factors for brain metastases [20–22].

6. Discussion and conclusion

One solution to the monotone likelihood problem, which is an important issue in Cox regression models, is Firth's penalized partial likelihood approach. However, the model selection criteria for this approach are yet to be studied, and heuristic criteria, AIC* and BIC*, are used in the SAS PHREG procedure. Therefore,

Table VI. The top five models based on AICF for the study data.

Model						AICF
	ntumor		*ptumor*		*neuro*	1753.84
	ntumor				*neuro*	1754.72
sex	*ntumor*		*ptumor*		*neuro*	1754.90
	ntumor	*diameter*	*ptumor*		*neuro*	1755.83
	ntumor		*ptumor*	*status*	*neuro*	1755.83

Note: *ntumor*, number of tumors; *diameter*, maximum diameter of largest tumor; *ptumor*, primary tumor category; *status*, extracerebral disease status; *neuro*, neurological symptoms; AICF, AIC for Firth's penalized partial likelihood approach.

Table VII. Parameter estimates of the best model based on AICF for the study data.

Covariate	*HR*	SE	95% CI		*p* (LR)
ntumor					
1 vs. 2–4	0.71	0.19	0.49	1.05	0.08
5–10 vs. 2–4	1.57	0.21	1.03	2.38	0.04
ptumor					
Breast vs. lung	0.94	0.26	0.57	1.57	0.82
GI vs. lung	1.60	0.36	0.79	3.26	0.21
Renal cell vs. lung	0.12	1.43	0.01	1.99	0.02
Others vs. lung	0.87	0.55	0.30	2.56	0.79
neuro (yes vs. no)	1.53	0.20	1.04	2.24	0.04

Note: *HR*, hazard ratio; *ntumor*, number of tumors; *ptumor*, primary tumor category; *neuro*, neurological symptoms; GI, gastrointestinal; LR, likelihood ratio; SE, standard error.

in this study, we proposed alternative criteria, AICF and BICF, which work for Firth's penalized partial likelihood approach. Moreover, we discussed the justification for adopting Firth's bias correction method in Cox regression models.

We showed that AICF, an estimator of the risk function based on KL information, does not include the penalty term of AIC*, $0.5\log|\mathbf{I}(\hat{\boldsymbol{\beta}})|$. Even in the penalized partial likelihood setting, non-penalized likelihood should be used for risk estimation. In addition, AICF is more efficient than AIC* in simulations and works well when addressing monotone likelihood. The simulation results revealed systematic bias in AIC*, and the model selection based on AICF showed superior predictive performance. In any case, AIC* is not recommended for model selection. An application using real data concluded that AICF has better properties than AIC* and that the latter leads to incorrect results. Moreover, we showed that BIC* has a negative penalty term in proportion to the number of regression parameters. The selection probability of the true model for BIC* was smaller than that of BICF for a large number of samples, indicating that it is not as efficient for selecting the true model. In summary, we showed that AICF or BICF is appropriate for model selection in monotone likelihood cases. AICF would be used for prediction, while BICF can be used for selecting the true model.

Because we obtained impressive results with alternative criteria, future studies should aim to examine other model evaluation criteria such as the *C*-index [23, 24]. Moreover, a similar problem occurs if one uses AIC and BIC based on the penalized log-likelihood under separation in logistic regression models.

Appendix A. Justification for using Firth's bias correction method in Cox regression models

Firth's bias correction method is based on the relationships $\kappa_{j,k}+\kappa_{jk}=0$, $\kappa_{j,k,l}+\kappa_{j,kl}+\kappa_{k,jl}+\kappa_{l,jk}+\kappa_{j,k,l}=0$, and $\kappa_{j,kl}=0$, as discussed in Section 3.2. However, the relationships $\kappa_{j,k,l}+\kappa_{j,kl}+\kappa_{k,jl}+\kappa_{l,jk}+\kappa_{j,k,l}=0$ and $\kappa_{j,kl}=0$ have not yet been evaluated in Cox regression models. Therefore, we only prove that $\kappa_{j,kl}=0$; it is well known that $\kappa_{j,k}+\kappa_{jk}=0$, and we can easily show that $\kappa_{j,k,l}+\kappa_{jkl}=0$.

Lemma 1 ($\kappa_{j,kl}=0$.)

Proof
If we insert $\boldsymbol{\beta}_0$ in the functions $U_j(\boldsymbol{\beta})$ and $U_{kl}(\boldsymbol{\beta})$,

$$U_j(\boldsymbol{\beta}_0)=\sum_{i=1}^{n}\int_0^1 H_{ij}(x)\,\mathrm{d}M_i(x),$$

$$H_{ij}(t)=Z_{ij}(t)-\frac{S_j^{(1)}(\boldsymbol{\beta}_0,t)}{S^{(0)}(\boldsymbol{\beta}_0,t)},$$

and

$$U_{kl}(\boldsymbol{\beta}_0)=\langle U_{kl}(\boldsymbol{\beta}_0)\rangle(1)+\sum_{i=1}^{n}\int_0^1 G_{kl}(x)\,\mathrm{d}M_i(x),$$

$$G_{kl}(t)=\frac{S_{kl}^{(2)}(\boldsymbol{\beta}_0,t)}{S^{(0)}(\boldsymbol{\beta}_0,t)}-\frac{S_k^{(1)}(\boldsymbol{\beta}_0,t)S_l^{(1)}(\boldsymbol{\beta}_0,t)}{\{S^{(0)}(\boldsymbol{\beta}_0,t)\}^2},$$

where $S_j^{(1)}(\boldsymbol{\beta},t)=n^{-1}\sum_{i=1}^{n}Y_i(t)Z_{ij}(t)\exp\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_i(t)\}$, $S_{kl}^{(2)}(\boldsymbol{\beta},t)=n^{-1}\sum_{i=1}^{n}Y_i(t)Z_{ik}(t)Z_{il}(t)\exp\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_i(t)\}$, and $\langle U_{kl}(\boldsymbol{\beta}_0)\rangle(t)=n\int_0^t G_{kl}(x)S^{(0)}(\boldsymbol{\beta}_0,x)h_0(x)\,\mathrm{d}x$.

We note that H_{ij} and G_{kl} are predictable processes according to the assumption made in Section 3.1. Here,

$$U_j(\boldsymbol{\beta}_0)U_{kl}(\boldsymbol{\beta}_0)=\langle U_{kl}(\boldsymbol{\beta}_0)\rangle(1)\sum_{i=1}^{n}\int_0^1 H_{ij}(x)\,\mathrm{d}M_i(x)+ \left\{\sum_{i=1}^{n}\int_0^1 H_{ij}(x)\,\mathrm{d}M_i(x)\right\}\left\{\sum_{i=1}^{n}\int_0^1 G_{kl}(x)\,\mathrm{d}M_i(x)\right\}.$$

From theorem 2.4.4 of [25], we have

$$\kappa_{j,kl} = n^{-1}\mathrm{E}\Bigg[\langle U_{kl}(\boldsymbol{\beta}_0)\rangle(1)\sum_{i=1}^{n}\int_0^1 H_{ij}(x)\,\mathrm{d}M_i(x)+ \\ \sum_{h=1}^{n}\sum_{i=1}^{n}\int_0^1 H_{hj}(x)G_{kl}(x)\,\mathrm{d}\langle M_h, M_i\rangle(x)\Bigg].$$

Because $M_i(t)$ is the counting process martingale,

$$\mathrm{E}\left[\sum_{i=1}^{n}\int_0^1 H_{ij}(x)\,\mathrm{d}M_i(x)\right] = 0,$$

and from the orthogonality of martingales

$$\langle M_h, M_i\rangle(t) = \begin{cases} \int_0^t Y_h(x)\exp\{\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{Z}_h(x)\}h_0(x)\,\mathrm{d}x, & h = i \\ 0, & h \neq i \end{cases}$$

It follows that

$$\begin{aligned}\kappa_{j,kl} &= 0 + n^{-1}\mathrm{E}\left[\int_0^1\left[\sum_{i=1}^{n}\left\{Z_{ij}(x) - \frac{S_j^{(1)}(\boldsymbol{\beta}_0, x)}{S^{(0)}(\boldsymbol{\beta}_0, x)}\right\}Y_i(x)\exp\{\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{Z}_i(x)\}\right]G_{kl}(x)h_0(x)\,\mathrm{d}x\right] \\ &= 0.\end{aligned}$$

Therefore, the relationships $\kappa_{j,k} + \kappa_{jk} = 0$, $\kappa_{j,k,l} + \kappa_{j,kl} + \kappa_{k,jl} + \kappa_{l,jk} + \kappa_{jkl} = 0$, and $\kappa_{j,kl} = 0$ are also true in Cox regression models. □

Appendix B. Consistency of $\hat{\boldsymbol{\beta}}$

In this section, we discuss the consistency of $\hat{\boldsymbol{\beta}}$. The following list of conditions will be assumed:

(A) $\int_0^1 h_0(x)\,\mathrm{d}x < \infty$.

(B) There exists a neighborhood $\mathcal{B}$ of $\boldsymbol{\beta}_0$ and a scalar function, $s^{(0)}$, a vector function, $\mathbf{s}^{(1)}$, and a matrix function, $\mathbf{s}^{(2)}$, defined on $\mathcal{B} \times [0, 1]$ such that for $k = 0, 1, 2$,

$$\sup_{t\in[0,1],\boldsymbol{\beta}\in\mathcal{B}} ||\mathbf{S}^{(k)}(\boldsymbol{\beta}, t) - \mathbf{s}^{(k)}(\boldsymbol{\beta}, t)|| \xrightarrow{P} 0.$$

(C) There exists $\delta > 0$ such that

$$n^{-1/2}\sup_{t\in[0,1],1\leq i\leq n} |\mathbf{Z}_i(t)|Y_i(t)1_{\{\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{Z}_i(t)>-\delta|\mathbf{Z}_i(t)|\}} \xrightarrow{P} 0,$$

where $1_{\{\}}$ is an indicator function.

(D) For all $\boldsymbol{\beta} \in \mathcal{B}$, $t \in [0, 1]$: $s^{(0)}(\boldsymbol{\beta}, t)$, $\mathbf{s}^{(1)}(\boldsymbol{\beta}, t) = \partial s^{(0)}(\boldsymbol{\beta}, t)/\partial\boldsymbol{\beta}$, and $\mathbf{s}^{(2)}(\boldsymbol{\beta}, t) = \partial\mathbf{s}^{(1)}(\boldsymbol{\beta}, t)/\partial\boldsymbol{\beta}^{\mathrm{T}}$ are continuous functions of $\boldsymbol{\beta} \in \mathcal{B}$, uniformly in $t \in [0, 1]$, $s^{(0)}$, $\mathbf{s}^{(1)}$, and $\mathbf{s}^{(2)}$ are bounded on $\mathcal{B} \times [0, 1]$; $s^{(0)}$ is bounded away from zero on $\mathcal{B} \times [0, 1]$, and the matrix $\boldsymbol{\Sigma}(\boldsymbol{\beta}_0)$ is positive definite.

These conditions are identical to those given by Andersen and Gill [12].

Lemma 2 ($\hat{\boldsymbol{\beta}} \to \boldsymbol{\beta}_0$.)

Proof
Consider the process

$$\begin{aligned}X_n^*(\boldsymbol{\beta}, 1) &= n^{-1}\{l^*(\mathbf{N}, \mathbf{Y}, \mathbf{Z}; \boldsymbol{\beta}) - l^*(\mathbf{N}, \mathbf{Y}, \mathbf{Z}; \boldsymbol{\beta}_0)\} \\ &= X_n(\boldsymbol{\beta}, 1) + 0.5n^{-1}\{\log|\mathbf{I}(\boldsymbol{\beta})| - \log|\mathbf{I}(\boldsymbol{\beta}_0)|\},\end{aligned}$$

where $X_n(\boldsymbol{\beta}, 1) = n^{-1}\{l(\mathbf{N}, \mathbf{Y}, \mathbf{Z}; \boldsymbol{\beta}) - l(\mathbf{N}, \mathbf{Y}, \mathbf{Z}; \boldsymbol{\beta}_0)\}$. To prove the consistency of the usual Cox regression estimator $\hat{\boldsymbol{\beta}}_{Cox}$, Andersen and Gill [12] showed that $X_n(\boldsymbol{\beta}, 1)$ is a concave function and proved that $X_n(\boldsymbol{\beta}, 1)$ converges in probability to

$$A(\boldsymbol{\beta}, 1) = \int_0^1 \left[(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^{\mathrm{T}} \mathbf{s}^{(1)}(\boldsymbol{\beta}_0, x) - \log \left\{ \frac{s^{(0)}(\boldsymbol{\beta}, x)}{s^{(0)}(\boldsymbol{\beta}_0, x)} \right\} s^{(0)}(\boldsymbol{\beta}_0, x) \right] h_0(t)\, \mathrm{d}x,$$

for each $\boldsymbol{\beta} \in \mathcal{B}$. The first derivative of $A(\boldsymbol{\beta}, 1)$ is zero at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, and the second derivative is minus a positive definite matrix. In other words, $A(\boldsymbol{\beta}, 1)$ is a concave function of $\boldsymbol{\beta}$ with a unique maximum at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. If $X_n(\boldsymbol{\beta}, 1)$ is a concave function, then $\hat{\boldsymbol{\beta}}_{Cox} \xrightarrow{P} \boldsymbol{\beta}_0$ by applying theorem II.1 of [12].

In the same manner, if $X_n^*(\boldsymbol{\beta}, 1)$ is concave and $X_n^*(\boldsymbol{\beta}, 1)$ converges to a concave function of $\boldsymbol{\beta}$ with a unique maximum at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, the consistency of $\hat{\boldsymbol{\beta}}$ can be shown by applying theorem II.1 of [12].

In monotone likelihood settings, the partial log-likelihood function converges to a finite value. Fixing $\beta_1, \beta_2, \dots, \beta_{j-1}, \beta_{j+1}, \dots, \beta_p$ to be $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_{j-1}, \hat{\beta}_{j+1}, \dots, \hat{\beta}_p$, for the parameter β_j, a real number c, and a constant d,

$$\exists c, \forall \beta_j \geq c, l(\mathbf{N}, \mathbf{Y}, \mathbf{Z}; \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_{j-1}, \beta_j, \hat{\beta}_{j+1}, \dots, \hat{\beta}_p) = d,$$

or

$$\exists c, \forall \beta_j \leq c, l(\mathbf{N}, \mathbf{Y}, \mathbf{Z}; \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_{j-1}, \beta_j, \hat{\beta}_{j+1}, \dots, \hat{\beta}_p) = d,$$

in monotone likelihood settings. Therefore, $X_n(\boldsymbol{\beta}, 1)$ is a concave function. Note that $l(\mathbf{N}, \mathbf{Y}, \mathbf{Z}; \boldsymbol{\beta}_0)$ and $\log |\mathbf{I}(\boldsymbol{\beta}_0)|$ are obviously finite constants. Thus, it is sufficient to show that $\log |\mathbf{I}(\boldsymbol{\beta})|$ is a concave function. Now, $\mathbf{I}(\boldsymbol{\beta})$ is a positive semidefinite matrix (see also Prentice [26]) because $l(\mathbf{N}, \mathbf{Y}, \mathbf{Z}; \boldsymbol{\beta})$ is a concave function. Generally, the function $\log |\mathbf{C}|$, where $|\mathbf{C}|$ is the determinant of a positive semidefinite matrix $\mathbf{C}$, is concave [27]. Therefore, $X_n^*(\boldsymbol{\beta}, 1)$ is a sum of concave functions.

We next show that $X_n^*(\boldsymbol{\beta}, 1)$ converges in probability to $A(\boldsymbol{\beta}, 1)$. According to the aforementioned result, $X_n(\boldsymbol{\beta}, 1)$ converges in probability to $A(\boldsymbol{\beta}, 1)$. Conditions (B) and (D) imply that, for each $\boldsymbol{\beta} \in \mathcal{B}$, $n^{-1}\mathbf{I}(\boldsymbol{\beta}) \xrightarrow{P} \boldsymbol{\Sigma}(\boldsymbol{\beta})$. Thus,

$$\begin{aligned} \log |\mathbf{I}(\boldsymbol{\beta})| - \log |\mathbf{I}(\boldsymbol{\beta}_0)| &= \log |n \cdot n^{-1}\mathbf{I}(\boldsymbol{\beta})| - \log |n \cdot n^{-1}\mathbf{I}(\boldsymbol{\beta}_0)| \\ &= \log |n^{-1}\mathbf{I}(\boldsymbol{\beta})| - \log |n^{-1}\mathbf{I}(\boldsymbol{\beta}_0)| \\ &= \log\{n^{-p}|\mathbf{I}(\boldsymbol{\beta})|\} - \log\{n^{-p}|\mathbf{I}(\boldsymbol{\beta}_0)|\} \end{aligned}$$

converges in probability to some finite quantity. Therefore,

$$0.5n^{-1} \left\{ \log |\mathbf{I}(\boldsymbol{\beta})| - \log |\mathbf{I}(\boldsymbol{\beta}_0)| \right\} \xrightarrow{P} 0,$$

and $X_n^*(\boldsymbol{\beta}, 1)$ converges in probability to $A(\boldsymbol{\beta}, 1)$, which is a concave function of $\boldsymbol{\beta}$ with a unique maximum at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$.

These facts establish that $\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}_0$ as $n \to \infty$.

Moreover, from this consistency, we can apply theorem 3.2 of [12]; therefore, $n^{-1}\mathbf{I}(\hat{\boldsymbol{\beta}}) \xrightarrow{P} \Sigma(\boldsymbol{\beta}_0)$ in monotone likelihood settings. □

Acknowledgements

The authors would like to thank the associate editor and two reviewers for very insightful and constructive comments that substantially improved the article. The authors would like to thank Professor Masaaki Yamamoto of Katsuta Hospital Mito Gamma House and Dr. Toru Serizawa of Tsukiji Neurological Clinic for providing the dataset of the JLGK0901 study. This work was supported by the Japan Society for the Promotion of Science KAKENHI grant numbers 26870099 and 16K16014 and the MHLW-funded Core Clinical Research Centers.

References

1. Cox DR. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* 1972; **34**(2):187–220.
2. Bryson MC, Johnson ME. The incidence of monotone likelihood in the Cox model. *Technometrics* 1981; **23**(4):381–383.

3. Heinze G, Schemper M. A solution to the problem of monotone likelihood in Cox regression. *Biometrics* 2001; **57**(1): 114–119.
4. Firth D. Bias reduction of maximum likelihood estimates. *Biometrika* 1993; **80**(1):27–38.
5. Heinze G. The application of Firth's procedure to Cox and logistic regression, Technical Report 10/1999, update in January 2001, Section of Clinical Biometrics, Department of Medical Computer Sciences University of Vienna, 2001.
6. Heinze G. A comparative investigation of methods for logistic regression with separated or nearly separated data. *Statistics in Medicine* 2006; **25**(24):4216–4226.
7. Heinze G, Schemper M. A solution to the problem of separation in logistic regression. *Statistics in Medicine* 2002; **21**(16):2409–2419.
8. Akaike H. Information theory and an extension of the maximum likelihood principle. In *In Second International Symposium on Information Theory*, Petrov BN, Csaki F (eds). Akademiai Kiado: Budapest, 1973; 267–281.
9. Schwarz G. Estimating the dimension of a model. *The Annals of Statistics* 1978; **6**(2):461–464.
10. Yamamoto M, Serizawa T, Shuto T, Akabane A, Higuchi Y, Kawagishi J, Yamanaka K, Sato Y, Jokura H, Yomo S, Nagano O, Kenai H, Moriki A, Suzuki S, Kida Y, Iwai Y, Hayashi M, Onishi H, Gondo M, Sato M, Akimitsu T, Kubo K, Kikuchi Y, Shibasaki T, Goto T, Takanashi M, Mori Y, Takakura K, Saeki N, Kunieda E, Aoyama H, Momoshima S, Tsuchiya K. Stereotactic radiosurgery for patients with multiple brain metastases (JLGK0901): a multi-institutional prospective observational study. *The Lancet Oncology* 2014; **15**(4):387–395.
11. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data* 2nd ed. John Wiley & Sons: New Jersey, 2002.
12. Andersen PK, Gill RD. Cox's regression model for counting processes: a large sample study. *The Annals of Statistics* 1982; **10**(4):1100–1120.
13. Cox DR, Snell EJ. A general definition of residuals. *Journal of the Royal Statistical Society, Series B* 1968; **30**(2):248–275.
14. Volinsky CT, Raftery AE. Bayesian information criterion for censored survival models. *Biometrics* 2000; **56**(1):256–262.
15. Xu R, Vaida F, Harrington DP. Using profile likelihood for semiparametric model selection with application to proportional hazards mixed models. *Statistica Sinica* 2009; **19**(2):819–842.
16. Claeskens G, Carroll RJ. An asymptotic theory for model selection inference in general semiparametric problems. *Biometrika* 2007; **94**(2):249–265.
17. Bailey KR. The asymptotic joint distribution of regression and survival parameter estimates in the Cox regression model. *The Annals of Statistics* 1983; **11**(1):39–48.
18. Johansen S. An extension of Cox's regression model. *International Statistical Review* 1983; **51**(2):165–174.
19. Murphy SA, van der Vaart AW. On profile likelihood. *Journal of the American Statistical Association* 2000; **95**(450): 449–465.
20. Joseph J, Adler J R, Cox R S, Hancock S L. Linear accelerator-based stereotaxic radiosurgery for brain metastases: the influence of number of lesions on survival. *Journal of Clinical Oncology* 1996; **14**(4):1085–1092.
21. Balm M, Hammack J. Leptomeningeal carcinomatosis. Presenting features and prognostic factors. *Archives of Neurology* 1996; **53**(7):626–632.
22. Sperduto PW, Kased N, Roberge D, Xu Z, Shanley R, Luo X, Sneed PK, Chao ST, Weil RJ, Suh J, Bhatt A, Jensen AW, Brown PD, Shih HA, Kirkpatrick J, Gaspar LE, Fiveash JB, Chiang V, Knisely JP, Sperduto CM, Lin N, Mehta M. Summary report on the graded prognostic assessment: an accurate and facile diagnosis-specific tool to estimate survival for patients with brain metastases. *Journal of Clinical Oncology* 2012; **30**(4):419–425.
23. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* 1996; **15**(4):361–387.
24. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the *C*-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine* 2011; **30**(10):1105–1117.
25. Fleming T R, Harrington D P. *Counting Processes and Survival Analysis*. John Wiley & Sons: New York, 1991.
26. Prentice RL, Self SG. Asymptotic distribution theory for Cox-type regression models with general relative risk form. *The Annals of Statistics* 1983; **11**(3):804–813.
27. Cover TM, Thomas JA. Determinant inequalities via information theory. *SIAM Journal on Matrix Analysis and Applications* 1988; **9**(3):384–392.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web-site.