Japanese Journal of Statistics and Data Science (2023) 6:337–359
https://doi.org/10.1007/s42081-023-00190-6

ORIGINAL PAPER

Statistics for Stochastic Processes

Asymptotic justification of maximum likelihood estimation for the proportional excess hazard model in analysis of cancer registry data

Sho Komukai[1] · Satoshi Hattori[1,2]

Received: 30 June 2022 / Revised: 13 January 2023 / Accepted: 28 January 2023 /
Published online: 10 March 2023

Abstract
Population-based cancer registry studies are conducted to investigate the various cancer question and have important impacts on cancer control. In order to investigate cancer prognosis from cancer registry data, it is necessary to adjust the effect of deaths from other causes, since cancer registry data include deaths from causes other than cancer. To correct for the effect of deaths from other causes, excess hazard models are often used. The concept of the excess hazard model is that the hazard function for any death in a cancer registry population is the sum of the hazard for cancer deaths, refer to the excess hazard, and the hazard for deaths from other causes. The Cox proportional hazard model for the excess hazard has been developed, and for this model, Perme et al. (Biostatistics 10:136–146, 2009) proposed the inference procedure of the regression coefficients using the techniques of the EM algorithm to compute the maximum likelihood estimator. In this article, we present the large sample properties for the maximum likelihood estimator. We introduce a consistent estimator of the variance for the regression coefficients based on the technique of the semiparametric theory and the consistency and the asymptotic normality of the estimator. The empirical property of variance estimator is investigated by the finite sample simulation studies. We also apply the variance estimator to cancer registry data for stomach, lung, and liver cancer patients from the Surveillance, Epidemiology, and End Results (SEER) database in U.S.

✉ Satoshi Hattori
hattoris@biostat.med.osaka-u.ac.jp

Sho Komukai
skomukai@biostat.med.osaka-u.ac.jp

1 Department of Biomedical Statistics, Graduate School of Medicine, Osaka University, Yamadaoka 2-2, Suita, Osaka 565-0871, Japan

2 Institute for Open and Transdisciplinary Research Initiatives, Osaka University, Yamadaoka 2-2, Suita, Osaka 565-0871, Japan

Keywords Cancer registry · Relative survival ratio · Net survival · Excess hazard model · Life table

1 Introduction

Cancer registries are effectively used in various cancer studies and play important roles in cancer control. A series of CONCORD studies addressed differences in cancer survival rates among nations for various cancer types, such as breast, colon, gastric, and prostate cancers (Coleman et al., 2008; Allemani et al., 2015, 2018). Derks et al. (2018) examined differences in survival outcomes due to differences in treatment policies among countries by the relative excess risks for older breast cancer patients in the Netherlands, Belgium, Ireland, England, and Greater Poland. These studies used data from cancer registries.

To address these scientific questions, rather than the overall survival, which is defined as the duration to the all-cause death, the cancer-specific survival is often of interest. Thus, the statistical analysis that accounts for the cause of death is appreciated. A potential approach is to apply methods for the competing risk analysis (Andersen et al.,1993, pp. 512–515; Fine and Gray, 1999). However, in cancer registries, reliable information on cause of death is hard to correct comprehensively. Then, in the field of cancer registry data analysis without the information on the cause of death, special survival analysis methods have been developed, in which the external data such as the life table of the general population are used to adjust the non-cancer death. This framework for inference of the cancer registry data is often referred as the relative survival framework (Perme et al., 2016; Kalager et al., 2021), since the relative survival ratio is one of key measure used in this field. The relative survival ratio is defined as the ratio of the overall survival to the non-cancer survival. Utilizing an external database of the life table for the non-cancer general population, various methods to handle the relative survival ratio has been proposed (Ederer et al., 1961; Hakulinen, 1982) and widely used in population studies (Coleman et al., 2008; Angle et al., 2014; Allemani et al., 2015). The net survival, which is defined as the survival probability if the cancer subject would not die due to reasons other than the cancer, is an alternative measure and is getting popular and popular after (Perme et al., 2012) introduced a novel estimator with sound theoretical justification. An application was reported by Allemani et al. (2018).

All these methods mentioned are for marginal quantities. Since cancer registry data consist of huge number of cancer patients, stratified analysis by age, gender, and so on with these simple methods is preferable in general without any strong statistical assumptions. On the other hand, regression analysis is also important. For example, for rare cancer types, the stratified analysis can be unstable. In assessing some covariates effects jointly, it would be useful to apply some regression models. Various regression models for cancer registry data were proposed including the parametric models (Rubio et al., 2018), the additive hazard model (Lambert et al., 2005; Cortese & Scheike, 2008) and the spline based nonproportional hazard model (Bolard et al., 2002; Gorgi et al., 2003). The Cox proportional hazard model, which is probably one of the most famous regression models for survival analysis, was also examined

(Hakulinen & Tenkanen, 1987; Estève et al., 1990; Sasieni, 1996; Dickman et al., 2004; Nelson et al., 2007; Perme et al., 2009). Sasieni (1996) introduced martingale estimating equations motivated by the partial likelihood. The unweighted estimating equation, which corresponds to the score function for the partial likelihood in the standard survival analysis, was not efficient for the Cox proportional hazard model for the net survival. Sasieni (1996) considered a weighted estimating equation, which gave an efficient estimator for the regression coefficients. However, to estimate the optimal weight, a smoothing technique was needed.

Perme et al. (2009) proposed the semiparametric maximum likelihood estimation. They successfully introduced a simple method to obtain the maximum likelihood estimator based on the expectation–maximization (EM) algorithm. The variance of the estimator was obtained with the Louis' method (Louis, 1982). Derks et al. (2018) applied this method to cancer registry data of older breast cancer patients in the Netherlands, Belgium, Ireland, England, and Greater Poland. Although the goodness of the method by Perme et al. (2009) was examined by their simulation study, asymptotic properties were not discussed. In this paper, we established asymptotic justification of the maximum likelihood estimator of the Cox proportional hazard model for the net survival by applying the general semiparametric efficiency theory. Instead of the Louis' variance estimator, we consider a variance estimator from the semiparametric theory.

The rest of the paper is organized as follows. In Sect. 2, we introduce cancer registry data and the EM-based inference procedure for the Cox proportional excess hazard model. In Sect. 3, we present the consistency of the maximum likelihood estimator for the regression coefficients. In Sect. 4, the asymptotic normality of the maximum likelihood estimator for the regression coefficients is presented. A consistent estimator of its asymptotic variance is also presented. In Sect. 5, we report the results of a simulation study, and in Sect. 6 we apply the proposed method to a real data from the Surveillance, Epidemiology, and End Results (SEER) Program. Some discussions are made in Sect. 7. All the theoretical details are placed in Appendixes.

2 Maximum likelihood estimation for Cox proportional excess hazard model

2.1 Notations and general settings for the cancer registry data

Analysis of cancer registry data generally requires two datasets: the cancer registry data and the population life tables. Cancer registry data consists of information on characteristics at diagnosis and the survival time for a subject diagnosed with cancer. Table 1 illustrates the data structure of the cancer registry data. Note that no information on the cause of death is included in the cancer registry data. The population life tables are a set of tables of annual mortality rates calculated by demographic variables for the general population, based on demographic statistics. Table 2 shows an example of the life table for the male population by age and calendar year. The information from the life table is used to correct the impact of death due to causes other than the cancer

Table 1 Example of cancer registry data: list of first five-observation of cancer registry data

ID	Time (Y)	Status	Age	Gender	Year	Stage
1	0.542	1	61	Male	1999	2
2	10.000	0	75	Male	1984	2
3	6.467	0	51	Male	1997	1
4	1.457	1	73	Male	1978	3
5	2.042	1	67	Male	1999	2

Table 2 Examples of the population life table for the male population with 60–64 years old in 1990–1994

		Year				
		1990	1991	1992	1993	1994
Age	60	1.146	1.140	1.140	1.128	1.087
	61	1.244	1.234	1.245	1.245	1.209
	62	1.341	1.330	1.350	1.368	1.331
	63	1.442	1.432	1.463	1.490	1.454
	64	1.555	1.541	1.582	1.609	1.574

The value of each cell implies 1-year mortality (%)

of interest. The mathematical formulations of the cancer registry data and the relative survival framework are given as follows.

Let Z be a bounded vector of baseline covariates in the cancer registry data. Typically, it consists of age at diagnosis, gender, year at diagnosis, and some other additional variables. Let T_O and C be the time-to-death due to any causes and the potential censoring time from the time of diagnosis. T_O may be censored by C. We suppose that $T = T_O \wedge C$ and $\Delta = I(T_O \leq C)$ are observed, where $A \wedge B$ is the minimum value of A and B and $I(\cdot)$ is the indicator function, which takes 1 if the event in bracket is true and 0 otherwise.

Let T_E and T_P be the potential time-to-death due to cancer and that due to reasons other than the cancer, respectively. Then, T_O is expressed as $T_O = T_E \wedge T_P$. Define $\Delta_E = I(T_E \leq T_P)$. We regard (T, Δ, Δ_E, Z) as the complete data, although the information of Δ_E is unobserved in the cancer registry data. The observed information is the triple (T, Δ, Z) for each subject in the cancer registry data. Let the corresponding counting process and the at-risk process denoted by $N(t) = I(T \leq t, \Delta = 1)$ and $Y(t) = I(T \geq t)$, respectively. Let τ be a constant satisfying $\Pr(T > \tau|Z) > 0$ for all Z. Suppose n i.i.d. copies of (T, Δ, Z) are observed and they are denoted by (T_i, Δ_i, Z_i). For other random variables, the subscript i is also used to represent the quantity for the ith subject.

Let $F_Z(z)$ be the distribution function for Z. The conditional survival function for T_O given Z is denoted by $S_O(t|Z) = \Pr(T_O > t|Z)$, and the corresponding hazard and cumulative hazard functions are denoted by $\lambda_O(t|Z)$ and $\Lambda_O(t|Z)$, respectively. These functions for T_E, T_P, and C are denoted in the same way with the subscripts E, P, and C, respectively. Suppose the assumption

$$\text{(A1) } T_E \perp T_P|Z$$

holds, where for any random variables A, B, and C, the conditional independence between A and B given C is denoted by $A \perp B|C$. Then, the hazard function for T_O is represented as the sum of those for T_E and T_P,

$$\lambda_O(t|Z) = \lambda_E(t|Z) + \lambda_P(t|Z).$$

The hazard function of T_E, $\lambda_E(t|Z)$, is called the excess hazard, representing the excess risk of death by cancer. The conditional hazard function $\lambda_P(t|Z)$ and the conditional survival function $S_P(t|Z)$ are calculated by an external database for population mortality and are regarded as known function.

2.2 Cox proportional excess hazard model

Suppose $\lambda_E(t|Z)$ is modeled via a Cox-type regression model

$$\lambda_E(t|Z) = \lambda(t)e^{\beta^T Z}, \tag{1}$$

where β is a vector of regression coefficients and $\lambda(t)$ is an unspecified baseline hazard function. Denote the baseline cumulative hazard function by $\Lambda(t) = \int_0^t \lambda(u)\mathrm{d}u$. Let β_0, $\lambda_0(t)$, and $\Lambda_0(t)$ be the true values of β, $\lambda(t)$, and $\Lambda(t)$, respectively. Furthermore, we assume

$$\text{(A2) } C \perp (T_E, T_P)\,|Z.$$

Under the assumptions of (A1) and (A2), the probability density function of the observed data (T, Δ, Z) is given by

$$\begin{aligned} &f_{T,\Delta,Z}(t, \delta, z; \Lambda, \beta) \\ &= \left\{\mathrm{d}\Lambda(t)e^{\beta^T z} + \mathrm{d}\Lambda_P(t|z)\right\}^{\delta} e^{-\Lambda(t)e^{\beta^T z} - \Lambda_P(t|z)} \mathrm{d}\Lambda_C(t|z)^{1-\delta} e^{-\Lambda_C(t|z)} \mathrm{d}F_Z(z), \end{aligned} \tag{2}$$

where $\mathrm{d}\Lambda(t) = \Lambda(t) - \Lambda(t-)$, $\mathrm{d}\Lambda_P(t|Z) = \Lambda_P(t|Z) - \Lambda_P(t-|Z)$, $\mathrm{d}\Lambda_C(t|z) = \Lambda_C(t|z) - \Lambda_C(t-|z)$, and $\mathrm{d}F_Z(z) = F_Z(z) - F_Z(z-)$. The observed likelihood function is

$$L(\Lambda, \beta) \propto \prod_{i=1}^{n} L(\Lambda, \beta; T_i, \Delta_i, Z_i), \tag{3}$$

where $L(\Lambda, \beta; T_i, \Delta_i, Z_i)$ is the contribution of the ith subject to the likelihood given by

$$L(\Lambda, \beta; T_i, \Delta_i, Z_i) = \left\{\mathrm{d}\Lambda(T_i)e^{\beta^T Z_i} + \mathrm{d}\Lambda_P(T_i|Z_i)\right\}^{\Delta_i} \exp\left\{-\Lambda(T_i)e^{\beta^T Z_i}\right\}. \tag{4}$$

Perme et al. (2009) proposed the semiparametric maximum likelihood estimator for the regression coefficients β, based on the EM algorithm. In constructing the semiparametric likelihood, $\Lambda(t)$ is regarded as a right-continuous and non-decreasing step function with $\Lambda(0) = 0$ and positive jump size $\lambda(t) > 0$ at all uncensored event time points to treat nonparametrically. The likelihood function for the complete data is

$$L_C(\Lambda, \beta) \propto \prod_{i=1}^{n} \left\{ \mathrm{d}\Lambda(T_i) e^{\beta^T Z_i} \right\}^{\Delta_{Ei}} \exp\left\{ -\Lambda(T_i) e^{\beta^T Z_i} \right\},$$

and the log-likelihood after profiling the baseline hazard function out is

$$\ell_{CP}(\beta) = \sum_{i=1}^{n} \left\{ \beta^T Z_i - \log \sum_{j=1}^{n} Y_j(T_i) e^{\beta^T Z_j} \right\} \Delta_{Ei}.$$

Set the initial values of λ and β as $\lambda^{(0)}$ and $\beta^{(0)}$, respectively. Then, the conditional expectation of $\ell_{CP}(\beta)$ given the observed data is

$$\begin{aligned} &Q(\beta; \lambda^{(0)}, \beta^{(0)}) \\ &= \sum_{i=1}^{n} \left\{ \beta^T Z_i - \log \sum_{j=1}^{n} Y_j(T_i) e^{\beta^T Z_j} \right\} \frac{\Delta_i \lambda^{(0)}(T_i) e^{\beta^{(0)T} Z_i}}{\lambda^{(0)}(T_i) e^{\beta^{(0)T} Z_i} + \lambda_P(T_i|Z_i)}. \end{aligned}$$

The value of β is updated by maximizing the Q function and the updated value is denoted by $\beta^{(1)}$. The value of λ is updated using the Breslow estimator as

$$\lambda^{(1)}(T_i) = \frac{\Delta_i \lambda^{(0)}(T_i) e^{\beta^{(0)T} Z_i}}{\lambda^{(0)}(T_i) e^{\beta^{(0)T} Z_i} + \lambda_P(T_i|Z_i)} \left\{ \sum_{j=1}^{n} Y_j(T_i) e^{\beta^{(1)T} Z_j} \right\}^{-1}.$$

By updating $\lambda^{(k)}$ and $\beta^{(k)}$ and repeating the computation and maximization of the Q-function, the estimators $\hat{\lambda}$ and $\hat{\beta}$ are obtained. The corresponding estimator of the baseline cumulative hazard function is represented by

$$\hat{\Lambda}(t) = \sum_{\{i: T_i \le t\}} \frac{\Delta_i \hat{\lambda}(T_i) e^{\hat{\beta}^T Z_i}}{\hat{\lambda}(T_i) e^{\hat{\beta}^T Z_i} + \lambda_P(T_i|Z_i)} \left\{ \sum_{j=1}^{n} Y_j(T_i) e^{\hat{\beta}^T Z_j} \right\}^{-1}. \tag{5}$$

3 Consistency

In this section, we prove the consistency of the maximum likelihood estimator. Suppose that β is in a compact set $\mathscr{B}$ and the covariance matrix of Z is positive definite. The existence of the pair of (Λ, β) which maximizes the observed likelihood function (3) is proved in Appendix A based on the techniques using in the proof of theorem 1 of Fang et al. (2005). The identifiability of (Λ, β), in the sense that $L(\Lambda, \beta; t, \delta, z) = L(\Lambda_0, \beta_0; t, \delta, z)$ implies $(\Lambda, \beta) = (\Lambda_0, \beta_0)$ on $t \in [0, \tau]$, is also shown in Appendix A.

The semiparametric model (2) has a set of the unknown parameters (β, η), where $\eta = \{\Lambda, \Lambda_C, F_Z\}$ is the nuisance parameter. Consider parametric submodels $\Lambda_{h_1}(t; \gamma_1) = \int_0^t \{1 + \gamma_1 h_1(u)\} \mathrm{d}\Lambda_0(u) = \int_0^t \{1 + \gamma_1 h_1(u)\} \lambda_0(u)\mathrm{d}u$, $\Lambda_{C,h_2}(t|Z; \gamma_2) = \int_0^t \{1 + \gamma_2 h_1(u, Z)\} \mathrm{d}\Lambda_C(u|Z) = \int_0^t \{1 + \gamma_2 h_2(u, Z)\} \lambda_C(u|Z)\mathrm{d}u$, and $F_{Z,h_3}(z; \gamma_3) = \int_0^t \{1 + \gamma_3 h_3(z)\} \mathrm{d}F_Z(z)$ where $h_1(u)$ and $h_2(u, Z)$ are an arbitrary function and $h_3(z)$ is a mean-zero measurable function with finite variance. The log-likelihood function based on (2) under the parametric submodel is defined by

$$
\begin{aligned}
\ell_n(\beta, \gamma; h) = & \sum_{i=1}^n \Delta_i \log\left[\{1 + \gamma_1 h_1(T_i)\} \mathrm{d}\Lambda_0(T_i) e^{\beta^T Z_i} + \mathrm{d}\Lambda_P(T_i|Z_i)\right] \\
& - \sum_{i=1}^n \int_0^{T_i} \{1 + \gamma_1 h_1(t)\} \mathrm{d}\Lambda_0(t) e^{\beta^T Z_i} \\
& + \sum_{i=1}^n (1 - \Delta_i) \log\left[\{1 + \gamma_2 h_2(T_i, Z_i)\} \mathrm{d}\Lambda_C(T_i|Z_i)\right] \\
& - \sum_{i=1}^n \int_0^{T_i} \{1 + \gamma_2 h_2(t, Z_i)\} \mathrm{d}\Lambda_C(t|Z_i) \\
& + \sum_{i=1}^n \{1 + \gamma_3 h_3(Z_i)\} \mathrm{d}F_Z(Z_i),
\end{aligned}
$$

where $\gamma = (\gamma_1, \gamma_2, \gamma_3)^T$ and $h = (h_1, h_2, h_3)^T$ Let

$$
W(t|Z; \beta, \Lambda) = \frac{\mathrm{d}\Lambda(t) e^{\beta^T Z}}{\mathrm{d}\Lambda(t) e^{\beta^T Z} + \mathrm{d}\Lambda_P(t|Z)}.
$$

Since the maximum likelihood estimator $\hat{\beta}$ maximizes the likelihood and then maximizes it under any parametric submodel, it satisfies

$$
U_n(\hat{\beta}; h) = \left(U_{n,\beta}(\hat{\beta}; h)^T, U_{n,\gamma}(\hat{\beta}; h)^T\right)^T = 0 \tag{6}
$$

for any h, where

$$
\begin{aligned}
U_{n,\beta}(\beta; h) = & \left.\frac{\partial}{\partial \beta} \ell_n(\beta, \gamma; h)\right|_{\gamma=0} \\
= & \sum_{i=1}^n \int_0^\tau Z_i W(t|Z_i; \beta, \Lambda_0) \left[\mathrm{d}N_i(t) - Y_i(t) \left\{d\Lambda_0(t) e^{\beta^T Z_i} + d\Lambda_P(t|Z_i)\right\}\right],
\end{aligned}
$$

$$
\begin{aligned}
U_{n,\gamma}(\beta;h) &= \left(U_{n,\gamma_1}(\beta;h_1), U_{n,\gamma_2}(\beta;h_2), U_{n,\gamma_3}(\beta;h_3)\right)^T,\\
U_{n,\gamma_1}(\beta;h_1) &= \frac{\partial}{\partial\gamma_1}\ell_n(\beta,\gamma;h)\Big|_{\gamma=0}\\
&= \sum_{i=1}^n \int_0^\tau h_1(t)W(t|Z_i;\beta,\Lambda_0)\left[\mathrm{d}N_i(t) - Y_i(t)\left\{d\Lambda_0(t)e^{\beta^T Z_i} + d\Lambda_P(t|Z_i)\right\}\right],\\
U_{n,\gamma_2}(\beta;h_2) &= \frac{\partial}{\partial\gamma_2}\ell_n(\beta,\gamma;h)\Big|_{\gamma=0} = \sum_{i=1}^n \int_0^\tau h_2(t,Z_i)\mathrm{d}M_{C,i}(t),\\
U_{n,\gamma_3}(\beta;h_3) &= \frac{\partial}{\partial\gamma_3}\ell_n(\beta,\gamma;h)\Big|_{\gamma=0} = \sum_{i=1}^n h_3(Z_i)\mathrm{d}F_Z(Z_i),
\end{aligned}
$$

and $M_C(t) = I(C \le t, \Delta = 0) - \int_0^t Y(u)\mathrm{d}\Lambda_C(u|Z)$ is a square integrable martingale with respect to some filtrations (Fleming & Harrington, 1991). Then, it can be shown that $E\left[U_1(\beta,\Lambda;h)\right] = 0$ for all bounded functions h on $t \in [0,\tau]$ and Z_1.

Theorem 1 *Under the assumptions (A1) and (A2), the maximum likelihood estimators are consistent; as $n \to \infty$, $\hat{\beta}$ converge in probability to β_0 and $\hat{\Lambda}(t)$ converge in probability to $\Lambda_0(t)$ uniformly in $t \in [0,\tau]$.*

Proof The estimator (5) is represented by

$$
\hat{\Lambda}(t) = \int_0^t \frac{1}{\sum_{j=1}^n Y_j(u)e^{\hat{\beta}^T Z_j}} \sum_{i=1}^n W(u|Z_i;\hat{\Lambda},\hat{\beta})\mathrm{d}N_i(u).
$$

Letting $h_1(t) = 1$ in the score equation (6) leads to this estimator. Since the vector of covariates Z is bounded and the parameter space $\mathscr{B}$ is compact, $e^{\beta^T Z}$ is bounded, and its upper bound is denoted by K_u. By the uniform low of large number (Pollard, 1990, page 41), $n^{-1}\sum_{j=1}^n Y_j(u)e^{\beta^T Z_j}$ converges almost surely to $E\left[Y(u)e^{\beta^T Z}\right] \in (0, K_u]$, uniformly in $t \in [0,\tau]$. By this result and $W(t|Z;\Lambda,\beta) \in [0,1]$ for all $t \in [0,\tau]$ and Z, $W(t|Z;\Lambda,\beta)$ and $n^{-1}\sum_{j=1}^n Y_j(u)e^{\hat{\beta}^T Z_j}$ are uniformly bounded on $[0,\tau]$. Then, we can use the procedures for proof of the consistency in Murphy et al. (1997). We give a sketch of the proof of consistency of $\hat{\beta}$ and $\hat{\Lambda}(t)$.

Define

$$
\tilde{\Lambda}(t) = \int_0^t \frac{1}{\sum_{j=1}^n Y_j(u)e^{\beta_0^T Z_j}} \sum_{i=1}^n W(u|Z_i;\Lambda_0,\beta_0)\mathrm{d}N_i(u).
$$

By the Lenglart inequality (Fleming and Harrington, 1991, page 113) and the uniform law of large numbers, we see that $\tilde{\Lambda}(t)$ converges almost surely to $\Lambda_0(t)$, uniformly in $t \in [0,\tau]$ as $n \to \infty$. Since $\hat{\Lambda}$ and $\hat{\beta}$ are the maximum likelihood estimator,

$$
n^{-1}\sum_{i=1}^n \left\{\log L(\hat{\Lambda},\hat{\beta};T_i,\Delta_i,Z_i) - \log L(\tilde{\Lambda},\beta_0;T_i,\Delta_i,Z_i)\right\} \ge 0,
$$

where $L(\Lambda, \beta; T_i, \Delta_i, Z_i)$ is defined in (4). Since $\hat{\Lambda}(t)$ and $\tilde{\Lambda}(t)$ are bounded, the ratios of their jump sizes are bounded and those ratios are of bounded variation as $n \to \infty$ in $t \in [0, \tau]$, we can use the results of the equation (A.5) in Murphy et al. (1997), and then those results imply that

$$E\left[\log L(\hat{\Lambda}, \hat{\beta}; T_i, \Delta_i, Z_i) - \log L(\tilde{\Lambda}, \beta_0; T_i, \Delta_i, Z_i)\right] \geq -o_P(1). \tag{7}$$

The function $\hat{\Lambda}(t)$ is non-decreasing and bounded function. By Helly's lemma (van der Vaart, 2000, page 9) and the compactness of $\mathscr{B}$, any subsequence indexed by n ($n = 1, 2, \cdots$) possesses a further subsequence satisfying $\hat{\beta} \to \beta^*$ for some β^* and $\hat{\Lambda}(t) \to \Lambda^*(t)$ for any $t \in [0, \tau]$ and some monotone function $\Lambda^*(t)$. Therefore, for any (t, δ, z),

$$\begin{aligned}&\log L(\hat{\Lambda}, \hat{\beta}; t, \delta, z)\\&\quad - \log L(\tilde{\Lambda}, \beta_0; t, \delta, z) \xrightarrow{P} \log L(\Lambda^*, \beta^*; t, \delta, z) - \log L(\Lambda_0, \beta_0; t, \delta, z).\end{aligned} \tag{8}$$

By the dominated convergence theorem, the expectation of the right-hand side of Eq. (8) under the true parameters Λ_0 and β_0, which is a minus of the Kullback–Leibler divergence, is nonpositive, and then by the result of the equation (7), it holds that

$$E\left[\log L(\Lambda^*, \beta^*; T, \Delta, Z) - \log L(\Lambda_0, \beta_0; T, \Delta, Z)\right] = 0.$$

By the identifiability of Λ and β and the lemma of (van der Vaart, 2000, page 62), we can conclude $\Lambda^* = \Lambda_0$ and $\beta^* = \beta_0$. Because any subsequence contains a further subsequence for which $\hat{\beta}$ and $\hat{\Lambda}$ converge uniformly to β_0 and Λ_0, respectively, their uniform convergence also holds for the entire sequence. $\square$

4 Asymptotic normality and variance estimation

In this section, the asymptotic normality of the maximum likelihood estimator is presented. To do so, we apply the semiparametric theory, and a consistent estimator of asymptotic variance is presented along the semiparametric theory.

Theorem 2 *Suppose that β_0 is in the interior of $\mathscr{B}$. Under the assumptions (A1) and (A2), $\sqrt{n}\left\{\hat{\beta} - \beta_0\right\}$ converge to a mean-zero Gaussian distribution with the variance $\Sigma_\beta(\beta_0, \Lambda_0; h^*)^{-1}$, where*

$$\begin{aligned}\Sigma_\beta(\beta, \Lambda; h^*) &= E\left[\left\{\int_0^\tau \left\{Z - h^*(t)\right\} W(t|Z; \beta, \Lambda)\mathrm{d}M(t)\right\}^{\otimes 2}\right],\\ h^*(t) &= \frac{E\left[W(t|Z; \beta_0, \Lambda_0)Y(t)Ze^{\beta_0^T Z}\right]}{E\left[W(t|Z; \beta_0, \Lambda_0)Y(t)e^{\beta_0^T Z}\right]},\end{aligned} \tag{9}$$

$M(t) = N(t) - \int_0^t Y(u)\left\{\mathrm{d}\Lambda_0(u)e^{\beta_0^T Z} + \mathrm{d}\Lambda_P(u|Z)\right\}$ *is a square integrable martingale with respect to some filtrations* (Fleming & Harrington, 1991)*, and* $V^{\otimes 2} = VV^T$ *for any column vector* V*. A consistent estimator of the asymptotic variance* (9) *is given by*

$$\hat{\Sigma}_\beta(\hat{\beta}, \hat{\Lambda}) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left\{Z_i - \frac{\sum_{k=1}^n W(t|Z_k; \hat{\beta}, \hat{\Lambda})Y_k(t)Z_k e^{\hat{\beta}^T Z_k}}{\sum_{j=1}^n W(t|Z_j; \hat{\beta}, \hat{\Lambda})Y_j(t)e^{\hat{\beta}^T Z_j}}\right\}^{\otimes 2} \times W(t|Z_i; \hat{\beta}, \hat{\Lambda})Y_i(t)e^{\hat{\beta}^T Z_i}\mathrm{d}\hat{\Lambda}(t). \tag{10}$$

Proof The nuisance tangent space for the nuisance parameter $\eta = \{\Lambda, \Lambda_C, F_C\}$ is given by a direct sum of three orthogonal linear spaces,

$$\Gamma = \Gamma_1 \oplus \Gamma_2 \oplus \Gamma_3,$$

where

$$\Gamma_1 = \left\{\int_0^\tau h_1(t)W(t|Z; \beta_0, \Lambda_0)\mathrm{d}M(t) \text{ for all function } h_1(t)\right\},$$
$$\Gamma_2 = \left\{\int_0^\tau h_2(t, Z)\mathrm{d}M_C(t) \text{ for all function } h_2(t, Z)\right\},$$
$$\Gamma_3 = \{h_3(Z) \text{ such that } E\left[h_3(Z)\right] = 0\}.$$

And the orthogonal complement of the nuisance tangent space Γ is written as

$$\Gamma^\perp = \left\{\int_0^\tau \left\{h_1(t, Z) - h_1^*(t)\right\} W(t|Z; \beta_0, \Lambda_0)\mathrm{d}M(t) \text{ for all function } h_1(t, Z)\right\},$$

where

$$h_1^*(t) = \frac{E\left[h_1(t, Z)W(t|Z; \beta_0, \Lambda_0)Y(t)e^{\beta_0^T Z}\right]}{E\left[W(t|Z; \beta_0, \Lambda_0)Y(t)e^{\beta_0^T Z}\right]}.$$

Details of the derivation of these nuisance tangent spaces and their orthogonal complements are given in Appendix B.

The efficient score function for β is constructed by orthogonal projection of the score function $U_{1,\beta}(\beta_0; h)$ onto the orthogonal space of Γ, and it is given by

$$U_{1,\beta}^{eff}(\beta_0; h^*) = \int_0^\tau \left\{Z_1 - h^*(t)\right\} W(t|Z_1; \beta_0, \Lambda_0)\mathrm{d}M_1(t),$$

where

$$h^*(t) = \frac{E\left[W(t|Z;\beta_0,\Lambda_0)Y(t)Ze^{\beta_0^T Z}\right]}{E\left[W(t|Z;\beta_0,\Lambda_0)Y(t)e^{\beta_0^T Z}\right]}.$$

Since the maximum likelihood estimator $\hat{\beta}$ satisfies $U_n(\hat{\beta}; h) = 0$, it is the solution to

$$\sum_{i=1}^{n}\int_0^{\tau}\{Z_i - h(t)\}\, W(t|Z_i;\beta,\Lambda_0)\left[\mathrm{d}N_i(t) - Y_i(t)\left\{\mathrm{d}\Lambda_0(t)e^{\beta^T Z_i} + d\Lambda_P(t|Z_i)\right\}\right]$$
$$= 0$$

with any bounded function h including h^*. The efficient influence function for ith subject is defined by

$$\psi_i(\beta_0,\Lambda_0;h^*) = \Sigma_\beta(\beta_0,\Lambda_0;h^*)^{-1}\int_0^{\tau}\left\{Z_i - h^*(t)\right\} W(t|Z_i;\beta_0,\Lambda_0)\mathrm{d}M_i(t), \tag{11}$$

where $\Sigma_\beta(\beta_0,\Lambda_0;h^*)$ is given as (9). Therefore, it holds that $\sqrt{n}\left(\hat{\beta} - \beta_0\right) = n^{-1/2}\sum_{i=1}^n \psi_i(\beta_0,\Lambda_0;h^*) + o_P(1)$ and it converges in law to the mean-zero Gaussian distribution with the variance function $\Sigma_\beta(\beta_0,\Lambda_0;h^*)^{-1}$.

The asymptotic variance (9) can be consistently estimated by replacing the theoretical quantities with the empirical ones. Then, a consistent estimator is given by (10). □

5 Simulation study

We conducted a simulation study to examine the behavior of the two variance estimators by (10) and Louis' method. The simulation settings were set by mimicking real cancer registry data and life tables. We considered four covariates, *age*, *gender*, *year*, and X. They were the age at diagnosis, the gender, the year of diagnosis, and a continuous variable. *Age*, *gender*, *year*, and X were generated from the normal distribution $N(60, 10^2)$, the Bernoulli distribution $B(1/2)$, the discrete uniform distribution $U(2000, 2010)$, and the standard normal distribution $N(0, 1)$, respectively. We generated T_E and T_P from the exponential distributions with hazard rate $\lambda_E(t|Z) = 0.20\exp\{\log 1.3\times\mathrm{st(age)} + \log 1.25\times\mathrm{gender} + \log 0.8\times\mathrm{st(year)} + \beta_X X\}$ and $\lambda_P(t|Z) = 0.02\exp\{\log 2.0\times\mathrm{st(age)} + \log 1.25\times\mathrm{gender} + \log 0.9\times\mathrm{st(year)}\}$, respectively, where $\mathrm{st(age)} = (\mathrm{age} - 60)/10$ and $\mathrm{st(year)} = (\mathrm{year} - 2000)/10$. We considered four scenarios on the magnitude of the association between T_E and X; $\beta_X = \log 1.0,\ \log 1.1,\ \log 1.2$, or $\log 1.3$ in Datasets 1-4, respectively. In all datasets, T_E and T_P were conditionally independent given the covariates Z. The potential censoring time C was generated from the uniform distribution on $(0, 30)$. We set the number of subjects n=200 or 1000. For each scenario, 1000 datasets were simulated.

We fitted the Cox model (1) with $Z = \{\mathrm{st(age)}, \mathrm{gender}, \mathrm{st(year)}, X\}$ in analyses. The regression coefficients were estimated by applying the maximum likelihood method with the EM algorithm by Perme et al. (2009), and the variance of those were estimated by the estimator (10) and the estimator from Louis's method. Because the survival function for the other cause death $S_P(t|Z)$ is regarded as a known function in the general cancer registry analyses, we used the true $S_P(t|Z)$ with $t = 1, 2, \ldots$ in the analyses. We matched three covariates age, gender, and year to extract $S_p(t|Z)$ for each cancer patient. We evaluated empirical mean of variance estimates, empirical power, and coverage probabilities (CP) for each regression coefficient.

The results for $n = 200$, 500, and 1000 cases are summarized in Tables 3, 4, and 5, respectively. The coverage probabilities of the proposed method (10) were close to the nominal level of 95% with $n = 500$ and $n = 1000$, whereas a little anti-conservativeness was observed with $n = 200$. The average and the empirical coverage probability for the variance estimates were almost identical between the method (10) and Louis's method throughout the simulation scenarios. It suggested that the two methods gave very similar estimates. To see that, we show the cross-plots the standard errors by the two methods in Fig. 1 for $n = 200$. For all the variables, the standard errors are laid near the diagonal line, indicating agreement between the two methods.

6 Illustration

We illustrate the proposed method with cancer registry data from the Surveillance, Epidemiology, and End Results (SEER) Program. We focused on a subgroup of all adult aged 60–69 years, who was diagnosed as stomach, lung, or liver cancers from 2005 to 2010 in 17 areas covering approximately 26.5% of the U.S. All patients were followed up to 10 years after diagnosis. The data were analyzed by cancer sites (stomach, lung, and liver). For each cancer site, the model (1) was applied with six covariates as explanatory variables; age at diagnosis, gender, year at diagnosis, race (White/Black/Others), stage (Localized/Regional/Distant), and income $<$ \$ 55,000, was applied. The regression coefficients were estimated by the EM-based method in Perme et al. (2009), and their variance were calculated by Louis's method or (10). To calculate $S_P(t|Z)$, we used the population life table of U.S., which is released from the SEER projects and it has information on annual survival by age, gender, year and race. For each cancer patient in the cancer registry, $S_P(t|Z)$ was extracted matching the four covariates of age, gender, year, and race.

In Table 6, we summarize patients' characteristics of the SEER database by cancer sites. For stomach, lung and liver cancers, 3987, 48,741 and 4608 patients were died among 5313, 56,412 and 5446 registered ones. The results of parameter estimation were summarized in Table 7. The two variance estimators gave very similar 95% CIs and p values.

Table 3 Simulation results with $n = 200$: rMSE implies root mean squared error ($\times 10^2$) and CP implies empirical coverage probability(%) for a 95% nominal level

						Semiparametric-based			Louis' method		
Dataset	Variable	True β	Mean of $\hat{\beta}$	Bias	rMSE	Mean of variance	Power	CP	Mean of variance	Power	CP
1	Age	0.2624	0.2621	−0.0003	9.812	0.835	82.0	92.5	0.835	82.0	92.5
	Gender	0.2231	0.2358	0.0127	18.433	3.021	28.3	93.8	3.027	28.2	93.8
	Year	− 0.2231	− 0.2362	− 0.0131	28.784	7.600	15.4	94.3	7.622	15.2	94.4
	X	0	− 0.0012	− 0.0012	18.247	3.003	6.1	93.9	3.009	6.0	94.0
2	Age	0.2624	0.2591	− 0.0032	9.558	0.818	82.0	93.2	0.819	81.9	93.0
	Gender	0.2231	0.2255	0.0023	17.906	2.965	27.4	94.8	2.972	27.2	94.9
	Year	− 0.2231	− 0.2298	− 0.0066	28.641	7.454	15.3	94.5	7.475	15.2	94.5
	X	0.0953	0.0903	− 0.005	17.913	2.950	9.3	93.8	2.956	9.2	93.8
3	Age	0.2624	0.2634	0.0011	9.2708	0.805	83.7	94.2	0.806	83.5	94.2
	Gender	0.2231	0.2334	0.0103	18.120	2.924	29.3	94.8	2.930	29.0	94.7
	Year	− 0.2231	− 0.2285	− 0.0054	28.042	7.314	14.8	95.3	7.332	14.9	95.4
	X	0.1823	0.1923	0.0099	18.362	2.919	22.1	92.9	2.925	21.9	93.0
4	Age	0.2624	0.2579	− 0.0045	9.424	0.793	81.7	94.5	0.793	81.7	94.3
	Gender	0.2231	0.2263	0.0031	18.236	2.881	28.6	93.0	2.887	28.5	93.0
	Year	− 0.2231	− 0.2116	0.0116	27.783	7.194	13.8	95.1	7.214	13.5	95.1
	X	0.2624	0.267	0.0047	17.966	2.888	35.5	92.8	2.893	35.3	92.9

Table 4 Simulation results with $n = 500$: rMSE implies root mean squared error ($\times 10^2$) and CP implies empirical coverage probability (%) for a 95% nominal level

						Semiparametric-based			Louis' method		
Dataset	Variable	True β	Mean of $\hat{\beta}$	Bias	rMSE	Mean of variance	Power	CP	Mean of variance	Power	CP
1	Age	0.2624	0.2564	− 0.0059	6.053	0.3216	99.2	93.9	0.3207	99.2	93.9
	Gender	0.2231	0.2283	0.0051	10.862	1.1791	56.0	95.0	1.1801	56.0	95.0
	Year	− 0.2231	− 0.2221	0.0011	17.600	2.9434	27.6	95.1	2.9472	27.6	95.1
	X	0	− 0.0012	− 0.0012	11.487	1.1723	6.1	93.9	1.1733	6.1	93.9
2	Age	0.2624	0.2571	− 0.0052	5.939	0.3153	99.3	93.8	0.3148	99.3	93.8
	Gender	0.2231	0.2281	0.0050	10.577	1.1564	56.8	95.1	1.1573	56.7	95.1
	Year	− 0.2231	− 0.2242	− 0.0011	17.329	2.8865	28.2	94.9	2.8900	28.3	94.9
	X	0.0953	0.0961	0.0008	11.290	1.1508	15.3	93.9	1.1518	15.3	94.1
3	Age	0.2624	0.2576	− 0.0048	5.935	0.3102	99.4	94.1	0.3097	99.4	94.0
	Gender	0.2231	0.2275	0.0043	10.570	1.138	56.6	94.8	1.1388	56.7	94.9
	Year	− 0.2231	− 0.2223	0.0008	17.204	2.8402	27.0	95.1	2.8435	27.0	95.1
	X	0.1823	0.1860	0.0037	11.239	1.1359	43.1	93.3	1.1368	43.1	93.3
4	Age	0.2624	0.2572	− 0.0052	5.861	0.306	99.2	94.3	0.3057	99.2	94.3
	Gender	0.2231	0.2281	0.0049	10.464	1.123	57.0	95.6	1.1239	56.9	95.6
	Year	− 0.2231	− 0.2225	0.0007	17.078	2.8027	27.1	94.5	2.8060	27.3	94.5
	X	0.2624	0.2675	0.0051	11.185	1.1259	71.0	93.4	1.1267	71.0	93.4

Table 5 Simulation results with $n = 1000$: rMSE implies root mean squared error ($\times 10^2$) and CP implies empirical coverage probability (%) for a 95% nominal level

						Semiparametric-based			Louis' method		
Dataset	Variable	True β	Mean of $\hat{\beta}$	Bias	rMSE	Mean of variance	Power	CP	Mean of variance	Power	CP
1	Age	0.2624	0.2554	– 0.0070	4.210	0.159	100	93.7	0.159	100	93.6
	Gender	0.2231	0.2218	– 0.0013	7.886	0.584	82.7	94.2	0.585	82.7	94.2
	Year	– 0.2231	– 0.2222	0.0009	12.368	1.460	43.9	95.6	1.461	43.9	95.6
	X	0	0.0033	0.0033	7.856	0.581	5.5	94.5	0.581	5.6	94.4
2	Age	0.2624	0.2560	– 0.0064	4.237	0.156	100	93.3	0.155	100	93.2
	Gender	0.2231	0.2233	0.0001	7.863	0.573	83.2	94.1	0.573	83.3	94.1
	Year	– 0.2231	– 0.2316	– 0.0084	12.746	1.431	50.5	93.9	1.431	50.3	93.8
	X	0.0953	0.0940	– 0.0013	7.938	0.570	25.4	94.0	0.570	25.4	94.0
3	Age	0.2624	0.2584	– 0.004	4.158	0.154	100	92.9	0.153	100	92.7
	Gender	0.2231	0.2250	0.0018	7.769	0.564	83.9	94.2	0.565	83.9	94.2
	Year	– 0.2231	– 0.2231	0.0000	12.305	1.411	46.8	94.8	1.412	46.7	94.8
	X	0.1823	0.1857	0.0034	7.877	0.564	70.1	94.1	0.564	70.1	94.1
4	Age	0.2624	0.2571	– 0.0053	4.026	0.151	100	94.1	0.151	100	94.1
	Gender	0.2231	0.2223	– 0.0009	7.273	0.556	85.2	95.9	0.556	85.2	95.9
	Year	– 0.2231	– 0.2267	– 0.0036	11.944	1.388	49.8	94.8	1.389	49.8	94.8
	X	0.2624	0.2695	0.0071	7.888	0.558	94.1	93.4	0.558	94.1	93.4

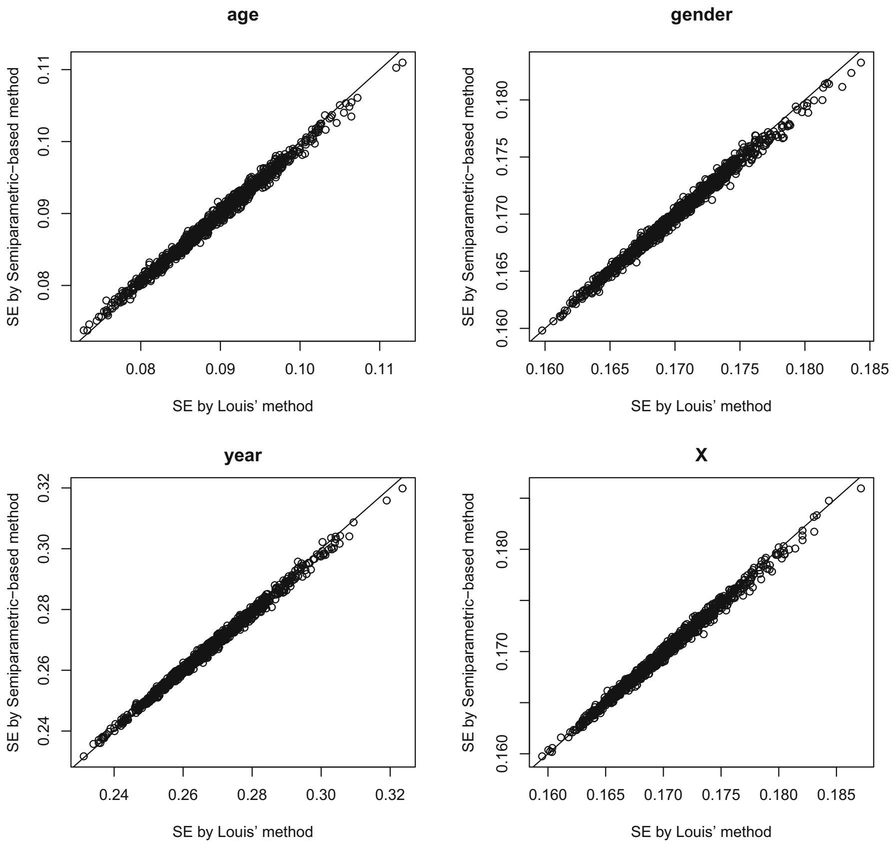

Fig. 1 Comparison of the standard error estimates between two methods for the 1000 simulated data in dataset 4 with $n = 200$; the horizontal line is for Louis' method and vertical line is for Semiparametric-based method

7 Discussion

Similarly to the standard survival analysis, the regression models play very important roles in analysis of cancer registry data. Many regression models were proposed in the relative survival setting (Rubio et al., 2018; Lambert et al., 2005; Cortese & Scheike, 2008; Bolard et al., 2002; Gorgi et al., 2003). With the substantial popularity of the original Cox proportional hazards model (Cox 1972), the Cox excess hazards regression would be one of the most important and appealing regression models in cancer registry data analysis. Successful introduction of a simple EM-based algorithm (Perme et al., 2009) for the maximum likelihood estimator is really appreciated and of practical value, and it was successfully applied in a real population study (Allemani et al., 2018). On the other hand, formal theoretical justification was left unclear. This paper contributes to fill the gap by showing consistency, asymptotic normality, and semiparametric efficiency. Although our theoretical justification covered only the

Table 6 Summary of SEER data by cancer sites (stomach, lung, and liver); the age at diagnosis, year at diagnosis, and the survival time were summarized by median with interquartile range (median [IQR]), and the other variables were summarized by the frequency and the proportion

	Stomach ($n = 5313$)	Lung ($n = 56{,}412$)	Liver ($n = 5446$)
Age at diagnosis	64 [62, 66]	64 [62, 66]	64 [61, 66]
Male (%)	3647 (68.6)	31,079 (55.1)	4154 (76.3)
Year at diagnosis	2007 [2006, 2008]	2007 [2006, 2008]	2007 [2006, 2009]
Race (%)			
White	3785 (71.2)	47,090 (83.5)	3756 (69.0)
Black	752 (14.2)	6296 (11.2)	669 (12.3)
Other	776 (14.6)	3026 (5.4)	1021 (18.7)
Stage (%)			
Localized	1454 (27.4)	10,620 (18.8)	2836 (52.1)
Regional	1830 (34.4)	13,846 (24.5)	1618 (29.7)
Distant	2029 (38.2)	31,946 (56.6)	992 (18.2)
Income $<$ \$ 55,000 (%)	1046 (19.7)	16299 (28.9)	932 (17.1)
Survival time	1.33 [0.42, 7.58]	0.92 [0.25, 3.17]	0.83 [0.17, 3.17]
Death (%)	3978 (74.9)	48741 (86.4)	4608 (84.6)

variance estimator (10), it also suggested the validity of Louis' estimator with the agreement between them observed in the simulation studies.

A typical way to use the regression model for cancer registry data is to evaluate conditional hazards given potential confounders as done by Derks et al. (2018); Schuil et al. (2018); Allemani et al. (2018). In recent years, studies combining cancer registry data with data from other databases have been conducted, and the search for factors that affect cancer prognosis has become increasingly important (Woods et al., 2021; Li et al., 2021). On the other hand, in making inference on marginal hazards, regression models also play very important roles. For example, Komukai and Hattori (2017, 2020) proposed doubly-robust inference procedures for the marginal net survival and relative survival ratio in the presence of covariate-dependent censoring, in which regression models for censoring time and the survival time were very crucial roles. Estimation of causal quantities under the relative survival setting was discussed based on the regression standardization by Syriopoulou et al. (2021). To incorporate the Cox excess hazards model in these settings, the sound theoretical basis of the model is very important. More specifically, the consistency and the efficient influence function (11) results for the estimators will be very useful theoretical results when showing the consistency and deriving the asymptotic variance of estimators incorporating the Cox excess hazards model, respectively. Our development would be helpful in developing rigorous methods for such incomplete data analysis of marginal quantities.

Finally, we conclude our paper by discussing the assumption (A1). It is a fundamental assumption in the analysis of cancer registry data, like the independent censoring assumption (Fleming and Harrington,1991, page 128) in the standard survival anal-

Table 7 Results for analyses of the SEER data by cancer sites (stomach, lung, and liver); the Cox proportional excess hazard model with covariates listed below as explanatory variables were applied by cancer sites

			Semiparametric-based		Louis' method	
Site	Variable	HR	95% CI	p value	95% CI	p value
Stomach	Age	1.035	1.019–1.051	< 0.001	1.018–1.051	< 0.001
	Male	1.011	0.925–1.104	0.812	0.923–1.106	0.816
	Year	0.963	0.936–0.992	0.011	0.935–0.992	0.013
	Race					
	Black	1.053	0.939–1.181	0.378	0.936–1.184	0.39
	Other	0.842	0.742–0.955	0.007	0.740–0.958	0.009
	Stage					
	Regional	2.226	1.919–2.582	< 0.001	1.911–2.592	< 0.001
	Distant	7.348	6.410–8.424	< 0.001	6.387–8.454	< 0.001
	Income	1.202	1.088–1.328	< 0.001	1.085–1.332	< 0.001
Lung	Age	1.009	1.006–1.013	< 0.001	1.006–1.013	< 0.001
	Male	1.225	1.203–1.246	< 0.001	1.202–1.248	< 0.001
	Year	1.002	0.996–1.008	0.584	0.995–1.008	0.609
	Race					
	Black	1.087	1.058–1.116	< 0.001	1.056–1.118	< 0.001
	Other	0.828	0.796–0.860	< 0.001	0.794–0.863	< 0.001
	Stage					
	Regional	1.992	1.935–2.052	< 0.001	1.926–2.061	< 0.001
	Distant	5.377	5.232–5.526	< 0.001	5.215–5.544	< 0.001
	Income	1.171	1.149–1.194	< 0.001	1.148–1.195	< 0.001
Liver	Age	1.01	0.996–1.024	0.150	0.996–1.025	0.158
	Male	1.134	1.038–1.238	0.005	1.036–1.241	0.006
	Year	0.957	0.933–0.981	< 0.001	0.932–0.982	< 0.001
	Race					
	Black	1.088	0.977–1.210	0.124	0.975–1.213	0.132
	Other	0.842	0.761–0.932	< 0.001	0.759–0.934	0.001
	Stage					
	Regional	2.210	2.029–2.408	< 0.001	2.025–2.413	< 0.001
	Distant	4.239	3.866–4.648	< 0.001	3.859–4.657	< 0.001
	Income	1.184	1.079–1.299	< 0.001	1.077–1.301	< 0.001

HR indicates the hazard ratio from the model (1)

ysis. To make the assumption (A1) satisfied, a simple idea is to collect and include many covariates so that (A1) holds. However, it also brings a difficulty specific to cancer registry data; even if additional covariates are collected in the cancer registries, the population life tables may not have them. This new missing data problem has been handled by Touraine et al. (2020) and Rubio et al. (2021). However, their development is not satisfactory, and further research is warranted possibly with an EM-based method like the proposed method in this paper.

Acknowledgements The first author's research was partly supported by Grant-in-Aid for Early-Career Scientists (20K19754) from the Ministry of Education, Science, Sports and Technology of Japan. The second author's research was partly supported by Grant-in-Aid for Challenging Exploratory Research (16K12403) and for Scientific Research (16H06299, 18H03208) from the Ministry of Education, Science, Sports and Technology of Japan. Computational calculations were performed at the Institute of Medical Science (the University of Tokyo).

Funding The research leading to these results received funding from the Ministry of Education, Science, Sports and Technology of Japan under Grant-in-Aid for Early-Career Scientists No. 20K19754, for Challenging Exploratory Research No. 16K12403, and for Scientific Research No. 16H06299 and No. 18H03208.

Declarations

Conflict of interest The authors declare no competing interests.

Appendix

Appendix A: Existence of the maximum likelihood estimator and identifiability of β_0 and Λ_0

The existence of the pair of the parameters (β, Λ) maximizing the observed likelihood (3) is proved by using the similar arguments to the proof of theorem 1 in Fang et al. (2005). In this Appendix, along this line, we prove the identifiability of (β_0, Λ_0) in the sense that $L(\Lambda, \beta; t, \delta, z) = L(\Lambda_0, \beta_0; t, \delta, z)$ implies $\beta = \beta_0$ and $\Lambda(t) = \Lambda_0$ for all $t \in [0, \tau]$.

Suppose the parameter space $\mathscr{B} \in R^p$ of β is compact, where p is the dimension of β. Since the vector of covariates Z is bounded, $e^{\beta^T Z}$ is also bounded, and its lower and upper bounds are denoted by K_l and K_u, respectively. Let $t_1 < t_2 < \cdots < t_k$ be the distinct failure times. Then, for any right-continuous and non-decreasing function $\Lambda(t)$, it holds that

$$\begin{aligned} 0 \le L(\beta, \Lambda) &\le \prod_{i=1}^{n} \{K_u \Lambda(T_i) + \Lambda_P(T_i|Z_i)\}^{\Delta_i} e^{-\Lambda(T_i)K_l} \\ &\le \prod_{i:T_i<t_k} \left\{\frac{K_u}{K_l} + \Lambda_P(T_i|Z_i)\right\}^{\Delta_i} \\ &\times \prod_{i:T_i=t_k} \left\{K_u \Lambda(t_k) e^{-\Lambda(t_k)K_l} + \Lambda_P(T_i|Z_i)\right\}^{\Delta_i}. \end{aligned} \tag{12}$$

Because forcing $\Lambda(T_i) = \Lambda(t_k)$ for all $T_i \ge t_k$ will increase the likelihood if t_k is sufficiently large value satisfying $\Lambda(t_k) \ge 1$, it suffices to restrict the space of $\Lambda(t)$ to the space Ω_0, where

$$\Omega_0 = \{\Lambda : \Lambda(t) \text{ is the right continuous and non-decreasing function with } \Lambda(t) = \Lambda(t_k) \text{ for all } t \geq t_k\}.$$

Let $A_M = \{\Lambda \in \Omega_0 : \Lambda(t_k) \leq M\}$ for any $0 < M < \infty$. Because $L(\beta, \Lambda)$ is continuous in β and Λ, it has a maximum in the compact subspace $\mathscr{B} \times A_M$ for any given M. Let $L^{(M)}$ be the maximum value of $L(\beta, \Lambda)$ in $\mathscr{B} \times A_M$. By $e^{-MK_l}M \to 0$ as $M \to \infty$, there exists an $M_0 \geq 1$ such that the right-hand side of (12) is less than $L^{(M_0)}$ for all Λ out of A_{M_0}. Therefore, the likelihood evaluated at any sequence Λ_m of Λ with $\Lambda_m(t_k)$ diverging to infinity as $m \to \infty$ will not approach the maximum value of $L(\beta, \Lambda)$. As a consequence, when maximizing the observed likelihood (3), we can restrict the compact subspace $\mathscr{B} \times A_{M_0}$. The existence of the maximum likelihood estimator can be proved by the continuity of the likelihood.

We prove that both of β_0 and Λ_0 are identifiable. By considering $L(\Lambda, \beta; t, 0, z) = L(\Lambda_0, \beta_0; t, 0, z)$, we have that $\Lambda(t)/\Lambda_0(t) = e^{-(\beta-\beta_0)^T Z}$ for all $t \leq \tau$ and Z such that $\Pr(T > \tau|Z) > 0$. Therefore, since $(\beta - \beta_0)^T Z$ is constant for all Z, it hold that $\beta = \beta_0$ if the covariance of Z is nondegenerate, and also we have $\Lambda(t) = \Lambda_0(t)$ for all $t \leq \tau$. By considering $L(\Lambda, \beta; t, 1, z) = L(\Lambda_0, \beta_0; t, 1, z)$, we also have $\mathrm{d}\Lambda(t) = \mathrm{d}\Lambda_0(t)$ for all $t \leq \tau$.

Appendix B: Nuisance tangent space and its orthogonal complement

Let $\mathcal{H}$ be a Hilbert space consisted of all p-dimensional measurable functions of (T, Δ, Z) with mean-zero and finite variance equipped with inner product $\langle h_1, h_2\rangle = E\left[h_1^T(T, \Delta, Z)h_2(T, \Delta, Z)\right]$. To derive the nuisance tangent space for the nuisance parameter $\eta = \{\Lambda, \Lambda_C, F_Z\}$, we consider parametric submodels $\Lambda_{h_1}(t; \gamma_1)$, $\Lambda_{C,h_2}(t|Z; \gamma_2)$, and $F_{Z,h_3}(z; \gamma_3)$ for Λ, Λ_C, and F_Z, respectively, which were defined in Sect. 3, where γ_1, γ_2, and γ_3 are the finite-dimensional nuisance parameters. Then, the nuisance tangent spaces for each nuisance parameter will be derived as the mean-square closure of all parametric submodel nuisance tangent spaces. Since the derivations of the nuisance tangent spaces Γ_2 and Γ_3 in Sect. 4, which are for the nuisance parameters Λ_C and F_Z, respectively, are the same as those of Section 5.2 in Tsiatis (2006), we only derive here the nuisance tangent space Γ_1, which is for the nuisance parameter Λ, in Theorem 1.

Again, we consider a parametric submodel $\Lambda_{h_1}(t; \gamma_1) = \int_0^t \{1 + \gamma_1 h_1(u)\}\, \mathrm{d}\Lambda_0(u) = \int_0^t \{1 + \gamma_1 h_1(u)\}\, \lambda_0(u)\mathrm{d}u$, where $h_1(u)$ is an arbitrary p-dimensional bounded function. The contribution to the log-likelihood function under the parametric submodel is

$$\ell_n(\beta, \gamma_1; h_1) = \sum_{i=1}^n \Delta_i \log\left[\{1 + \gamma_1 h_1(T_i)\}\, \mathrm{d}\Lambda_0(T_i)e^{\beta^T Z_i} + \mathrm{d}\Lambda_P(T_i|Z_i)\right] - \sum_{i=1}^n \int_0^{T_i} \{1 + \gamma_1 h_1(t)\}\, \mathrm{d}\Lambda_0(t)e^{\beta^T Z_i}.$$

Taking derivatives of $\ell_n(\beta, \gamma_1; h_1)$ with respect to γ_1, and evaluating $\beta = \beta_0$ and $\gamma_1 = 0$, we obtain the score function

$$U_{n,\gamma_1}(\beta_0; h) = \sum_{i=1}^{n} \int_0^{\tau} h_1(t) W(t|Z_i; \beta_0, \Lambda_0) \mathrm{d}M_i(t).$$

Then, the score function for this parametric submodel is in the nuisance tangent space Γ_1. Since any element of $\mathcal{H}$ can be approximated by a sequence of bounded function (Tsiatis 2006, Section 4), the score function with parametric submodel without the boundedness of $h_1(t)$ is also in Γ_1.

For any parametric submodel $\Lambda(t; \gamma_1) = \int_0^t \lambda(u; \gamma_1)\mathrm{d}u$, the score function with respect to γ_1, setting $\gamma_1 = 0$ and $\beta = \beta_0$, is expressed as

$$U_{1,\gamma_1}(\beta_0) = \int_0^{\tau} \left\{ \frac{\partial}{\partial \gamma_1} \log \lambda(t; \gamma_1) \bigg|_{\gamma_1 = 0} \right\} W(t|Z; \beta_0, \Lambda_0) \mathrm{d}M(t).$$

Then, this score function is in the nuisance tangent space Γ_1. On the other hand, we can demonstrate that the score function for the some parametric submodel included in Γ_1, such as $\Lambda_{h_1}(t; \gamma_1) = \int_0^t \{1 + \gamma_1 h_1(u)\} \mathrm{d}\Lambda_0(u)$, is an element of a parametric submodel nuisance tangent space. Therefore, it holds that the nuisance tangent space for $\Lambda(t)$ is equal to Γ_1.

$\Gamma_1 \perp \Gamma_2$ can be easily proved under the assumption $(A2)$ and $\Gamma_i \perp \Gamma_3$ $(i = 1, 2)$ can be also proved by $E[\alpha_i^T h_3(Z)] = 0$, where $\alpha_i \in \Gamma_i$ $(i = 1, 2)$ and $h_3(Z) \in \Gamma_3$. Then the nuisance tangent space for the nuisance parameter $\eta = \{\Lambda, \Lambda_C, F_Z\}$ is given by the direct sum of three orthogonal spaces, $\Gamma = \Gamma_1 \oplus \Gamma_2 \oplus \Gamma_3$. The orthogonal complement $\Gamma^{\perp}$ is obtained by applying the almost same procedures as the proof of Theorem 5.5 in Tsiatis (2006).

References

Allemani, C., Weir, H. K., Carreira, H., Harewood, R., Spika, D., Wang, X. S., et al. (2015). Global surveillance of cancer survival 1995–2009: Analysis of individual data for 25,676,887 patients from 279 population-based registries in 67 countries (CONCORD-2). *Lancet, 385*, 977–1010.

Allemani, C., Matsuda, T., Carlo, V. D., Harewood, R., Matz, M., Maja, N., et al. (2018). Global surveillance of trends in cancer survival 2000–14 (CONCORD-3): Analysis of individual records for 37,513,025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. *Lancet, 39*, 1023–75.

Andersen, P. K., Borgan, O., Gill, R. D., & Keiding, N. (1993). *Statistical models based on counting processes*. Springer.

Bolard, P., Quantin, C., Abrahamowicz, M., Estève, J., Giorgi, R., Chadha-Boreham, H., et al. (2002). Assessing time-by-covariate interactions in relative survival models using restrictive cubic spline functions. *Journal of Cancer Epidemiology and Prevention, 7*, 113–122.

Coleman, M. P., Quaresma, Q., Berrino, F., Lutz, J., Angelis, R. D., Capocaccia, R., et al. (2008). Cancer survival in five continents: A worldwide population-based study (CONCORD). *Lancet Oncology, 9*, 730–756.

Cortese, G., & Scheike, T. H. (2008). Dynamic regression hazards models for relative survival. *Statistics in Medicine, 27*, 3563–3584.

Derks, M. G. M., Bastiaannet, E., Kiderlen, M., Hilling, D. E., Boelens, P. G., Walsh, P. M., et al. (2018). Variation in treatment and survival of older patients with nonmetastatic breast cancer in five European countries: A population-based cohort study from the EURECCA Breast Cancer Group. *British Journal of Cancer, 119*, 121–129.

Dickman, P. W., Sloggett, A., Hills, M., & Hakulinen, T. (2004). Regression models for relative survival. *Statistics in Medicine, 23*(1), 51–64.

Ederer, F., Axitell, L. M., & Cutler, S. J. (1961). The relative survival rate: A statistical methodology. *National Cancer Institute Monograph, 6*, 101–121.

Estève, J., Benhamou, E., Croasdale, M., & Raymond, L. (1990). Relative survival and the estimation of net survival: Elements for further discussion. *Statistics in Medicine, 9*, 529–538.

Fang, H.-B., Li, G., & Sun, J. (2005). Maximum likelihood estimation in a semiparametric logistic/proportional-hazard mixture model. *Scandinavian Journal of Statistics, 32*, 59–75.

Fine, J. P., & Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association, 94*, 496–509.

Fleming, T. R., & Harrington, D. P. (1991). *Counting processes and survival analysis*. Wiley.

Gorgi, R., Abrahamowicz, M., Quantin, C., Bolard, P., Estève, J., Gouvernet, J., & Faivre, J. (2003). A relative survival regression model using B-spline functions to model non-proportional hazards. *Statistics in Medicine, 22*, 2767–2784.

Hakulinen, T. (1982). Cancer survival corrected for heterogeneity in patient withdrawal. *Biometrics, 38*, 933–942.

Hakulinen, T., & Tenkanen, L. (1987). Regression analysis of relative survival rates. *Journal of the Royal Statistical Society, Series C., 36*, 309–317.

Kalager, M., Adami, H.-O., Lagergren, P., Steindorf, K., & Dickman, P. W. (2021). Cancer outcomes research-a European challenge: Measures of the cancer burden. *Molecular Oncology, 15*, 3223–3241.

Komukai, S., & Hattori, S. (2017). Doubly robust estimator for net survival rate in analyses of cancer registry data. *Biometrics, 73*, 124–133.

Komukai, S., & Hattori, S. (2020). Doubly robust inference procedure for relative survival ratio in population-based cancer registry data. *Statistics in Medicine, 39*(13), 1884–1900.

Lambert, P. C., Smith, L. K., Jones, D. R., & Botha, J. L. (2005). Additive and multiplicative covariate regression models for relative survival incorporating fractional polynomials for time-dependent effects. *Statistics in Medicine, 24*, 3871–3885.

Li, M., Reintals, M., D'Onise, K., Farshid, G., Holmes, A., Joshi, R., Karapetis, C. S., Miller, C. L., Olver, I. N., Buckley, E. S., Townsend, A., Walters, D., & Roder, D. M. (2021). Investigating the breast cancer screening-treatment-mortality pathway of women diagnosed with invasive breast cancer: Results from linked health data. *European Journal of Cancer Care, 31*, e13539. https://doi.org/10.1111/ecc.13539

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B, 44*, 226–233.

Murphy, S. A., Rossini, A. J., & van der Vaart, A. W. (1997). Maximum likelihood estimation in the proportional odds model. *Journal of the American Statistical Association, 92*, 968–976.

Nelson, C. P., Lambert, P. C., Squire, I. B., & Jones, D. R. (2007). Flexible parametric models for relative survival, with application in coronary heart disease. *Statistics in Medicine, 26*, 5486–5498.

Perme, M. P., Stare, J., & Estève, J. (2012). On estimation in relative survival. *Biometrics, 68*, 113–120.

Perme, M. P., Henderson, R., & Stare, J. (2009). An approach to estimation in relative survival regression. *Biostatistics, 10*, 136–146.

Perme, M. P., Estève, J., & Rachet, B. (2016). Analysing population-based cancer survival–settling the controversies. *BMC Cancer, 16*, 933. https://doi.org/10.1186/s12885-016-2967-9

Pollard, D. (1990). *Empirical processes: Theory and applications*. Institute of Mathematical Statistics.

Rubio F. J., Remontet L., Jewell N. P. & Belot A. (2018). On a general structure for hazard-based regression models: an application to population-based cancer research. *Statistical Methods in Medical Research*.

Rubio, F. J., Rachet, B., Giorgi, R., Maringe, C., Belot, A., & the CENSUR working survival group,. (2021). On models for the estimation of the excess mortality hazard in case of insufficiently stratified life tables. *Biostatistics, 22*(1), 51–67.

Sasieni, P. D. (1996). Proportional excess hazards. *Biometrika, 83*, 127–141.

Schuil, H., Derks, M., Liefers, G.-J., Portielje, J., van de Velde, C., Syed, B., Green, A., Ellis, I., Cheung, K.-L., & Bastiaannet, E. (2018). Treatment strategies and survival outcomes in older women with breast cancer: A comparative study between the FOCUS cohort and Nottingham cohort. *Journal of Geriatric Oncology, 9*, 635–641.

Syriopoulou, E., Rutherford, M. R., & Lambert, P. C. (2021). Inverse probability weighting and doubly robust standardization in the relative survival framework. *Statistics in Medicine, 40*, 6069–6092.
Touraine, C., Graféo, N., Giorgi, R., & the CENSUR working survival group. (2020). More accurate cancer-related excess mortality through correcting background mortality for extra variables. *Statistical Methods in Medical Research, 29*(1), 122–136.
Tsiatis, A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer.
Van der Vaart, A. W. (2000). *Asymptotic statistics* (Vol. 3). Cambridge University Press.
Woods, L. M., Rachet, B., Morris, M., Bhaskaran, K., & Coleman, M. P. (2021). Are socio-economic inequalities in breast cancer survival explained by peri-diagnostic factors? *BMC Cancer, 21*, 485. https://doi.org/10.1186/s12885-021-08087-x