Received: 3 July 2020 | Revised: 7 November 2020 | Accepted: 21 January 2021

DOI: 10.1002/bimj.202000209

Biometrical Journal

RESEARCH PAPER

Cox regression analysis for distorted covariates with an unknown distortion function

Yanyan Liu[1] | **Yuanshan Wu**[2] | **Jing Zhang**[2] | **Haibo Zhou**[3]

[1] School of Mathematics and Statistics, Wuhan University, Wuhan, Hubei, P. R. China

[2] School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan, Hubei, P. R. China

[3] Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Correspondence
Jing Zhang, School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan, Hubei 430073, China; Haibo Zhou, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7420, USA.
Email: jing66@zuel.edu.cn; zhou@bios.unc.edu

Funding information
U.S. National Institute of Health, Grant/Award Numbers: P42ES031007, P30ES010126; National Natural Science Foundation of China, Grant/Award Numbers: 11771366, 11971362, 11671311, 11901581, 12071483

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

Abstract

We study inference for censored survival data where some covariates are distorted by some unknown functions of an observable confounding variable in a multiplicative form. An example of this kind of data in medical studies is normalizing some important observed exposure variables by patients' body mass index , weight, or age. Such a phenomenon also appears frequently in environmental studies where an ambient measure is used for normalization and in genomic studies where the library size needs to be normalized for the next generation sequencing of data. We propose a new covariate-adjusted Cox proportional hazards regression model and utilize the kernel smoothing method to estimate the distorting function, then employ an estimated maximum likelihood method to derive the estimator for the regression parameters. We establish the large sample properties of the proposed estimator. Extensive simulation studies demonstrate that the proposed estimator performs well in correcting the bias arising from distortion. A real dataset from the National Wilms' Tumor Study is used to illustrate the proposed approach.

KEYWORDS
covariate adjustment, Cox regression model, distorting function, estimated maximum likelihood method, multiplicative effect

1 | INTRODUCTION

In real studies, the primary covariates sometimes are not directly recorded in their true values, but rather they are observed in a distorted form, where the distortion is in the form of a multiplicative factor. These types of data do not get sufficient attention as other types of covariate measurement error problems, even though they are also quite wide prevalent in real studies. For example, when releasing household data on energy use, in order to maintain confidentiality, the U.S. Department of Energy multiplied the survey data by some randomly selected numbers before publication (Hwang, 1986). Therefore, the contaminated data available to the public are $\widetilde{X} = X \cdot U$, where X and U, respectively, denote the true data and the randomly selected number. This multiplicative contamination structure is also very common in biomedical studies, in the form of normalization, as some primary covariates are often normalized by a confounder such as body mass index (BMI = weight/height^2) or by other measures of body configuration or age. For instance, in a study of the relationship between the fibrinogen level (FIB) and serum transferrin level (TRF) among hemodialysis patients, Kaysen et al. (2002) found that BMI has a great influence on FIB and TRF and may distort the true relationship between them. Therefore, they proposed a calibration method where they divide the observed FIB and TRF by the confounding variable BMI. This implies a multiplicative structure between the unobserved primary variables and the confounding variable. Such a phenomenon also appears frequently in environmental studies where the ambient measure is used for normalization and in genomic studies where the library size needs to be normalized for the next generation sequencing of data.

In some situations, however, the precise nature of the multiplicative relationship between the primary variables and the confounding variable could be unknown, and in this case the naive practice of dividing by the confounding variable may result in biased estimators or losing the power of statistical inference. To overcome these difficulties, Sentürk and Müller (2005) considered a more flexible multiplicative form which is an unknown function of the confounding variable U. They proposed a covariate adjustment method for the linear regression model, where both the response (Y) and the covariates (X) are distorted by an observable confounder U, that is, $\widetilde{X} = \phi(U)X$, $\widetilde{Y} = \varphi(U)Y$, where $\widetilde{X}$ and $\widetilde{Y}$ are observable distorted covariates and response, $\phi(\cdot)$ and $\varphi(\cdot)$ are unknown smooth distorting functions. Directly applying the widely used ordinary least squares method to the contaminated data, $(\widetilde{X}, \widetilde{Y})$ will result in biased and inconsistent estimates. Sentürk and Müller (2005) corrected the bias by linking it to a varying-coefficient regression model, then utilized the bin method (Fan & Zhang, 2000) to obtain consistent estimators (Sentürk & Müller, 2006). Related research includes Nguyen and Sentürk (2008) on generalizing this method to the case of multiple distorting covariates, Sentürk and Müller (2009) on extending it to a generalized linear model, and Zhang et al. (2012, 2013) on the nonlinear regression model and the partial linear model, respectively. More recently, Cui et al. (2009) developed a direct plug-in estimation procedure for a nonlinear regression model with one confounding variable. They proposed to estimate the distorting functions $\varphi(\cdot)$ and $\phi(\cdot)$ by nonparametrically regressing the response and predictors on the distorting variable and obtained the estimates $(\widehat{X}, \widehat{Y})$ for the unobservable response and predictors, then conducted the nonlinear least-squares method on the estimated counterparts $(\widehat{X}, \widehat{Y})$. Zhang et al. (2012) further applied this direct plug-in method to a semiparametric model by incorporating dimension reduction techniques. To relax the parametric assumptions and some restrictive conditions on distorting functions in the existing literature, Delaigle et al. (2016) proposed a more flexible nonparametric estimator for the regression function.

In this paper, we focus on investigating censored survival data where the response of interest is a right-censored survival time and the primary predictor X is distorted by an observable confounding variable U through the multiplicative form $\widetilde{X} = \phi(U)X$, where $\phi(\cdot)$ is the unknown distorted function. A reasonable identifiability condition for this structure is $E\{\phi(U)\} = 1$ corresponding to the assumption that the mean distorting effect vanishes (Sentürk & Müller, 2005). The existing methods mentioned earlier cannot be applicable here due to censoring. Furthermore, the existing methods for censored survival data with mismeasured covariates (e.g., Prentice, 1982; Wang et al., 1997; Zhou & Pepe, 1995; Zhou & Wang, 2000; Huang & Wang, 2000; Hu & Lin, 2002) cannot handle this multiplicative distortion. To make a valid inference, we propose a covariate-adjusted Cox proportional hazards regression to address this multiplicative contamination structure. Inspired by Cui et al. (2009), we first employ the nonparametric regression to obtain the consistent estimator of the distorting function $\phi(\cdot)$ through the kernel smoothing method,and then obtain the estimates for the true covariate X by $\hat{X} = \tilde{X}/\hat{\phi}(U)$. Then, the regression parameters are estimated by maximizing the partial likelihood on the estimated data. The proposed approach has several distinctive advantages. First, the contamination structure we considered is more general, which includes a large class of confounding mechanisms, for example, $\phi(\cdot) = 1$ means there is no contamination and $\phi(U) = U$ represents the contamination structure $\widetilde{X} = X \cdot U$. So the applicability of our proposed method can be quite broad. Second, the computation of this method is simple and fast, which will greatly facilitate its implementation in real application.

The rest of the article is organized as follows. In Section 2, we introduce the covariate-adjusted Cox regression for the multiplicative contaminated data and present the proposed covariate-calibration method. In Section 3, we establish the asymptotic properties of the proposed estimates. In Section 4, we present the simulation results to evaluate the finite sample performance of the proposed estimates. In Section 5, we apply the proposed method to a dataset from the National Wilms' Tumor Study (NWTS). Some concluding remarks are given in Section 6. The details of all technical proofs and the additional simulation results are presented in the Supporting Information.

2 | COX REGRESSION WITH THE MULTIPLICATIVE CONTAMINATION STRUCTURE

2.1 | Model, data, and contamination structure

To fix notation, let T denote the survival time, C denote the censoring time, $\widetilde{T} = \min(T, C)$ denote the observed time, and $\Delta = I(T \leq C)$ denote the failure indicator. Let $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_p)^{\mathrm{T}}$ and X be the associated covariates where X is the one that subjects to multiplicative contamination. Assume that the censoring mechanism is random, that is, the survival time T and the censoring time C are conditionally independent given $\mathbf{Z}$ and X. The proportional hazards regression model (Cox, 1972) assumes that the conditional hazard function of the survival time T associated with covariates $\mathbf{Z}$ and X takes the form of

$$\lambda(t|\mathbf{Z}, X) = \lambda_0(t)\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z} + \gamma X),$$

where $\lambda_0(t)$ is the baseline hazard function, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^{\mathrm{T}}$ and γ are the unknown regression coefficients. We assume the scalar covariate X is not observed precisely, while the p-dimensional covariate $\mathbf{Z}$ could be accurately observed. Assume that the observed data consist of n subjects, denoted by $(\widetilde{T}_i, \Delta_i, \mathbf{Z}_i, U_i, \widetilde{X}_i)$, $i = 1, \ldots, n$, which are independent samples from $(\widetilde{T}, \Delta, \mathbf{Z}, U, \widetilde{X})$. Instead of exact X_i, we observe $\widetilde{X}_i$ such that

$$\widetilde{X}_i = \phi(U_i)X_i, \tag{1}$$

where U_i is an observable variable and independent of X_i and $\phi(\cdot)$ is an unknown link function. To make the model identifiable, we assume that $E\{\phi(U_i)\} = 1$, which implies that the distorting effect vanishes on average.

We aim to infer the regression parameters $\boldsymbol{\beta}$ and γ based on the observations available. When X_i are observed without contamination, maximizing the partial likelihood function (Cox, 1975)

$$L_n(\boldsymbol{\beta}, \gamma) = \prod_{i=1}^{n} \left\{ \frac{\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_i + \gamma X_i)}{\sum_{j=1}^{n} I(\widetilde{T}_j > \widetilde{T}_i)\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_j + \gamma X_j)} \right\}^{\Delta_i} \tag{2}$$

can offer the estimates for $\boldsymbol{\beta}$ and γ. It is evident that (2) cannot be used when X_i are unobservable or contaminated.

Note that the established methods of the Cox regression with additive contamination structure $\widetilde{X} = X + U$ always require error U to be independent of X (e.g., Huang & Wang, 2000; Li & Ryan, 2004). The direct application of the additive error structure methods to the current setting is not feasible. To illustrate this, even though the multiplicative contamination structure (1) can also be rewritten as an additive structure,

$$\widetilde{X} = X + X\{\phi(U) - 1\}, \tag{3}$$

or

$$\log\widetilde{X} = \log X + \log\{\phi(U)\}, \tag{4}$$

the error $X\{\phi(U) - 1\}$ is not independent of X; hence, the methods mentioned above cannot be applicable here. If one takes the logarithmic transformation assuming the related quantities are positive, then one would arrive at the additive covariate contamination structure (4). Here the error term $\log\{\phi(U)\}$ is independent of $\log X$, but extra variation needs to be accounted for in the back-transformation procedure. Moreover, the routine approximately corrected score method for the Cox regression at the scale $\log X$ would result in the biased estimate when the Cox regression is against X.

2.2 | Covariate-calibration method

Our proposed approach is based on directly calibrating X_i. Note that

$$\phi(u) = \frac{E(\widetilde{X}|U = u)}{E(X)} = \frac{E(\widetilde{X}|U = u)}{E(\widetilde{X})}.$$

We can employ the commonly used the Nadaraya–Watson kernel smoothing estimate for $\psi(u) = E(\widetilde{X}|U = u)$, which is given by

$$\widehat{\psi}(u) = \frac{\sum_{i=1}^{n} K\{(u - U_i)/h_n\}\widetilde{X}_i}{\sum_{i=1}^{n} K\{(u - U_i)/h_n\}},$$

where $K(\cdot)$ is the kernel smoothing function and h_n is the bandwidth. Since $\bar{\widetilde{X}}_n = n^{-1}\sum_{i=1}^{n}\widetilde{X}_i$ converges to $E(\widetilde{X})$ almost surely by using the strong law of large numbers, we can obtain a consistent estimate for $\phi(u)$ as $\widehat{\phi}(u) = \widehat{\psi}(u)/\bar{\widetilde{X}}_n$. Following (1), we propose a calibration of X_i by $\widehat{X}_i = \widetilde{X}_i/\widehat{\phi}(U_i)$. Therefore, we can construct an estimated partial likelihood function using $\widehat{X}_i$ as follows:

$$\widehat{L}_n(\boldsymbol{\beta}, \gamma) = \prod_{i=1}^{n}\left\{\frac{\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_i + \gamma\widehat{X}_i)}{\sum_{j=1}^{n} I(\widetilde{T}_j > \widetilde{T}_i)\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_j + \gamma\widehat{X}_j)}\right\}^{\Delta_i}. \tag{5}$$

The proposed estimator $(\widehat{\boldsymbol{\beta}}, \widehat{\gamma})$ was defined as the maximizer for $\widehat{L}_n(\boldsymbol{\beta}, \gamma)$, that is,

$$(\widehat{\boldsymbol{\beta}}, \widehat{\gamma}) = \operatorname{argmax}_{(\boldsymbol{\beta},\gamma)}\widehat{L}_n(\boldsymbol{\beta}, \gamma). \tag{6}$$

2.3 | Bandwidth selection

In real data analysis, it is desirable to have an automatically data-driven method for selecting the bandwidth parameter h_n. Here we employ the cross-validation (CV) method to choose the optimal h_n. In particular, let $p(u)$ denote the density function of U, then the kernel estimate of $p(u)$ is denoted as

$$\widehat{p}(u) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{u - U_i}{h_n}\right).$$

Following Rudemo (1982) and Bowman (1984), we define an integrated squared error (ISE) as follows:

$$\begin{aligned}\mathrm{ISE}(h_n) &= \int \{\widehat{p}(u) - p(u)\}^2 \mathrm{d}u \\ &= \int \{\widehat{p}(u)\}^2 \mathrm{d}u - 2\int \widehat{p}(u)p(u)\mathrm{d}u + \int \{p(u)\}^2 \mathrm{d}u.\end{aligned} \tag{7}$$

As the third term of (7) is free of h_n, the minimizer of $\mathrm{ISE}(h_n)$ is the same as the minimizer of the sum of the first two terms of (7). Let $\widehat{p}_{(-i)}(\cdot)$ be the leave-one-out kernel density estimator, that is,

$$\widehat{p}_{(-i)}(u) = \frac{1}{nh_n}\sum_{j\neq i}^{n} K\left(\frac{u - U_j}{h_n}\right).$$

The second term of (7) can be consistently estimated by $-2n^{-1}\sum_{i=1}^{n}\widehat{p}_{(-i)}(U_i)$. Therefore, we propose a CV criterion as follows:

$$\mathrm{CV}(h_n) = \int \{\widehat{p}(u)\}^2 \mathrm{d}u - 2n^{-1}\sum_{i=1}^{n}\widehat{p}_{(-i)}(U_i).$$

Denote

$$\widehat{h}_{n,\mathrm{opt}} = \operatorname{argmin}_{h_n}\mathrm{CV}(h_n),$$

which is considered as the optimal bandwidth parameter.

3 | ASYMPTOTIC PROPERTIES

We set $\theta = (\boldsymbol{\beta}^{\mathrm{T}}, \gamma)^{\mathrm{T}}$, let $\hat{\theta} = (\hat{\boldsymbol{\beta}}^{\mathrm{T}}, \hat{\gamma})^{\mathrm{T}}$ and $\theta_0 = (\boldsymbol{\beta}_0^{\mathrm{T}}, \gamma_0)^{\mathrm{T}}$, respectively, represent the estimation and the true value of the regression parameter θ. The following theorem gives the consistency and asymptotic normality of the proposed estimator $\hat{\theta}$ when $n \to \infty$. The regularity conditions and the proofs of this theorem are given in Appendixes A and B, respectively.

Theorem 1. *Let $\hat{\theta} = (\hat{\boldsymbol{\beta}}^{\mathrm{T}}, \hat{\gamma})^{\mathrm{T}}$ be defined by (6). If conditions C1–C9 in Appendix A are satisfied, the following results hold:*

(i) $\hat{\theta}$ converges in probability to the true value θ_0,
(ii) $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(\mathbf{0}, \Sigma^{-1}(\Sigma + \Omega)\Sigma^{-1})$,

where $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$ is defined in condition C4, Σ_{11} denotes the pth-order sequential principal minor of Σ, and $\zeta = (-\Sigma_{12}^{\mathrm{T}}\gamma_0, -\Sigma_{22}\gamma_0)^{\mathrm{T}}$, $\Omega = \frac{Var(\tilde{X}) - Var(X)}{\{E(X)\}^2}\zeta\zeta^{\mathrm{T}}$.

The above theorem establishes the asymptotic normality of the proposed estimator $\hat{\theta}$. Furthermore, we can obtain the asymptotic distribution of $\hat{\boldsymbol{\beta}}$ and $\hat{\gamma}$, respectively. In particular, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathrm{N}(\mathbf{0}, (\Sigma^{-1})_p)$, $\sqrt{n}(\hat{\gamma} - \gamma_0) \xrightarrow{d} \mathrm{N}(0, (\Sigma^{-1})_{(p+1,p+1)} + \frac{\mathrm{Var}(\tilde{X}) - \mathrm{Var}(X)}{\{E(X)\}^2}\gamma_0^2)$, where $(\Sigma^{-1})_p$ denotes the pth-order sequential principal minor of matrix Σ^{-1} and $(\Sigma^{-1})_{(p+1,p+1)}$ denotes its $(p+1)$th diagonal element. We give a few remarks on the asymptotic covariance matrix. If there is no distortion with $\phi(\cdot) = 1$, we can estimate θ by maximizing the partial likelihood (2), the asymptotic covariance matrix of $\hat{\theta}$ is Σ^{-1}. So the term $\Sigma^{-1}\Omega\Sigma^{-1}$ is caused by the distortion. Furthermore, the limiting variance for $\hat{\gamma}$ includes some unknown components which need to be estimated; therefore, we can use the sandwich method and the plug-in estimation to obtain the standard error and construct the confidence region for $\hat{\gamma}$.

4 | SIMULATION STUDIES

We conduct extensive simulations to investigate the finite-sample performance of the proposed estimator $(\hat{\boldsymbol{\beta}}, \hat{\gamma})$ and compare it with the other two estimators. The first one is the naive estimator $(\hat{\boldsymbol{\beta}}_N, \hat{\gamma}_N)$, which ignores the contamination and directly uses $\tilde{X}$ to replace X, and the second one is the oracle estimator $(\hat{\boldsymbol{\beta}}_O, \hat{\gamma}_O)$, which is obtained by assuming that X is known.

The survival times T_i are generated from the Cox proportional hazards model with the conditional hazard function given by

$$\lambda(t|\mathbf{Z}_i, X_i) = \lambda_0(t)\exp(\boldsymbol{\beta}_0{}^{\mathrm{T}}\mathbf{Z}_i + \gamma_0 X_i),$$

where $\lambda_0(t) = 1$, $\boldsymbol{\beta}_0 = (1, 0.5)^{\mathrm{T}}$, and $\gamma_0 = 1.5$. We generate the covariate $\mathbf{Z}_i = (Z_{i1}, Z_{i2})^{\mathrm{T}}$ from a multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = (0.8^{|j-k|})$ $(j, k = 1, 2)$, X_i from $N(1, 0.5^2)$ and the confounding covariate U_i from a uniform distribution Unif(2,6). The censoring time $C = \tilde{C} \wedge \tau$, where $\tilde{C}$ is generated from Unif$(0, \tau + 2)$ and the study duration τ is chosen to yield the desirable censoring rates. Here we consider two forms of distortion function $\phi(u) = (u + 3)/7$ and $\phi(u) = 3(u + 1)^2/79$, which satisfy $E\{\phi(U_i)\} = 1$. To estimate the distorting function, we choose the Gaussian kernel function $K(t) = \exp(-t^2/2)/\sqrt{2\pi}$ and adopt the leave-one-out CV method to select the bandwidth. We consider the sample size $n = 100$ and $n = 200$, coupled with the censoring rates (CR) of 20%, 40%, and 80%. For each configuration, we repeat 1,000 simulations.

Tables 1 and 2 summarize the simulation results of $(\hat{\boldsymbol{\beta}}, \hat{\gamma})$, $(\hat{\boldsymbol{\beta}}_N, \hat{\gamma}_N)$, and $(\hat{\boldsymbol{\beta}}_O, \hat{\gamma}_O)$ under different distortion functions and different censoring rates for sample sizes $n = 100$ and $n = 200$, respectively. We make the following observations: (i) As expected, in terms of the mean-square error or the coverage probability, the oracle estimator $\hat{\gamma}_O$ and our proposed estimator $\hat{\gamma}$ are all superior to the naive estimator $\hat{\gamma}_N$. Not surprisingly, the naive estimator $\hat{\gamma}_N$ is seriously biased. For example, under the censoring rate of 20% and $\phi(u) = 3(u + 1)^2/79$ in Table 1, the bias for $\hat{\gamma}_N$ is -0.810, more than half of its real value 1.5, while the bias for proposed estimator $\hat{\gamma}$ is only -0.047; moreover, the coverage probability for $\hat{\gamma}_N$ is 0.006, almost equals to zero. It seems from the simulation results of $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\beta}}_N$, and $\hat{\boldsymbol{\beta}}_O$ that these three methods give similar performance on the

TABLE 1 Simulation results for $\boldsymbol{\beta}$ and γ under sample size $n = 100$

			$\phi(u) = (u+3)/7$					$\phi(u) = 3(u+1)^2/79$				
CR	**Method**	**Para.**	**Bias**	**SD**	**SE**	**MSE**	**CP**	**Bias**	**SD**	**SE**	**MSE**	**CP**
20%	Proposed	$\hat{\beta}_1$	0.020	0.224	0.222	0.051	0.960	0.006	0.225	0.221	0.051	0.946
		$\hat{\beta}_2$	0.007	0.211	0.206	0.044	0.942	0.001	0.212	0.206	0.045	0.943
		$\hat{\gamma}$	0.021	0.286	0.278	0.082	0.946	−0.047	0.292	0.267	0.087	0.906
	Naive	$\hat{\beta}_{1N}$	−0.006	0.224	0.221	0.050	0.945	−0.089	0.223	0.217	0.058	0.911
		$\hat{\beta}_{2N}$	−0.007	0.211	0.206	0.044	0.948	−0.045	0.211	0.205	0.047	0.933
		$\hat{\gamma}_N$	−0.217	0.254	0.235	0.112	0.780	−0.810	0.171	0.160	0.686	0.006
	Oracle	$\hat{\beta}_{1O}$	0.030	0.222	0.222	0.050	0.953	0.030	0.222	0.222	0.050	0.953
		$\hat{\beta}_{2O}$	0.011	0.210	0.206	0.044	0.936	0.011	0.210	0.206	0.044	0.936
		$\hat{\gamma}_O$	0.044	0.276	0.279	0.078	0.952	0.044	0.276	0.279	0.078	0.952
40%	Proposed	$\hat{\beta}_1$	0.024	0.259	0.253	0.068	0.944	0.011	0.260	0.252	0.068	0.941
		$\hat{\beta}_2$	0.012	0.251	0.237	0.063	0.932	0.007	0.251	0.237	0.063	0.939
		$\hat{\gamma}$	0.031	0.334	0.316	0.113	0.937	−0.042	0.341	0.303	0.118	0.897
	Naive	$\hat{\beta}_{1N}$	0.001	0.258	0.251	0.067	0.946	−0.073	0.256	0.247	0.071	0.920
		$\hat{\beta}_{2N}$	0.000	0.250	0.237	0.063	0.942	−0.034	0.250	0.236	0.063	0.933
		$\hat{\gamma}_N$	−0.220	0.299	0.265	0.137	0.791	−0.824	0.200	0.180	0.718	0.026
	Oracle	$\hat{\beta}_{1O}$	0.034	0.257	0.253	0.067	0.942	0.034	0.257	0.253	0.067	0.942
		$\hat{\beta}_{2O}$	0.015	0.250	0.237	0.063	0.928	0.015	0.250	0.237	0.063	0.928
		$\hat{\gamma}_O$	0.050	0.324	0.315	0.107	0.951	0.050	0.324	0.315	0.107	0.951
80%	Proposed	$\hat{\beta}_1$	0.072	0.454	0.436	0.212	0.932	0.065	0.461	0.435	0.217	0.931
		$\hat{\beta}_2$	0.031	0.454	0.413	0.207	0.914	0.028	0.452	0.412	0.205	0.912
		$\hat{\gamma}$	0.107	0.589	0.542	0.359	0.939	0.020	0.572	0.517	0.328	0.927
	Naive	$\hat{\beta}_{1N}$	0.059	0.466	0.434	0.221	0.930	0.012	0.460	0.428	0.212	0.927
		$\hat{\beta}_{2N}$	0.022	0.450	0.413	0.203	0.913	0.002	0.449	0.410	0.201	0.915
		$\hat{\gamma}_N$	−0.183	0.507	0.449	0.291	0.883	−0.830	0.340	0.304	0.805	0.254
	Oracle	$\hat{\beta}_{1O}$	0.080	0.460	0.437	0.218	0.935	0.080	0.460	0.437	0.218	0.935
		$\hat{\beta}_{2O}$	0.033	0.458	0.414	0.211	0.909	0.033	0.458	0.414	0.211	0.909
		$\hat{\gamma}_O$	0.125	0.592	0.539	0.366	0.939	0.125	0.592	0.539	0.366	0.939

The true value of the parameters $\beta_1 = 1$, $\beta_2 = 0.5$, $\gamma = 1.5$; $\phi(\cdot)$, the distortion function; CR, the censoring rate; Bias, the estimate value minus the true value; SD, the standard deviation; SE, the estimate of SD; MSE, the mean-square error; CP, empirical coverage percentage of the 95% confidence interval.

estimation of $\boldsymbol{\beta}$ which correspondes to $\mathbf{Z}$. This phenomenon is reasonable, as the covariates $\mathbf{Z}$ can be observed accurately; the three methods all use the accurate information of covariates $\mathbf{Z}$ when estimating their corresponding parameters. (ii) The proposed estimator $(\hat{\boldsymbol{\beta}}, \hat{\gamma})$ is essentially unbiased and comparable with the oracle estimator under different settings, even for the cases with a high censoring rate of 80%. For example, in the case of the censoring rate= 40% and $\phi(u) = 3(u+1)^2/79$ in Table 1, the relative efficiency $SD(\hat{\gamma})/SD(\hat{\gamma}_O) = 0.341/0.324 = 1.05$, very close to 1. (iii) Our proposed method performs stably with the choice of the distortion function, while the naive method performs worse if we chose $\phi(u) = 3(u+1)^2/79$. The coverage probabilities of $\hat{\gamma}_N$ for $\phi(u) = 3(u+1)^2/79$ are almost equal or close to zero. These simulation results demonstrate that the proposed covariate-calibration approach can effectively overcome the negative effect arising from the covariate contamination and meanwhile exhibits good performance.

Furthermore, we consider the informative censoring mechanism where the censoring time C is generated from $\mathrm{Unif}(0, \mathrm{c} \cdot |\mathrm{Z}_1 - \mathrm{Z}_2|)$, c is chosen to achieve the desirable censoring rates of 20%, 40%, and 80%. The remaining setups are kept the same as before. The simulation results are summarized in Tables 3 and 4, from which we can see the proposed method also performs well when the completely random censoring assumption does not hold.

To evaluate the performance of the proposed estimation procedure when the contamination structure is mis-specified, we consider the additive contamination structure $\widetilde{X} = X + U$, where X is generated from $N(2, 0.5^2)$ and the confounding covariate U is generated from $N(0, 0.3^2)$, the remaining setups are kept the same as before. The simulation results are

TABLE 2 Simulation results for $\boldsymbol{\beta}$ and γ under sample size $n = 200$

			$\phi(u) = (u+3)/7$					$\phi(u) = 3(u+1)^2/79$				
CR	**Method**	**Para.**	**Bias**	**SD**	**SE**	**MSE**	**CP**	**Bias**	**SD**	**SE**	**MSE**	**CP**
20%	Proposed	$\hat{\beta}_1$	0.009	0.156	0.151	0.024	0.939	0.004	0.157	0.151	0.025	0.936
		$\hat{\beta}_2$	0.004	0.151	0.142	0.023	0.936	0.002	0.152	0.142	0.023	0.935
		$\hat{\gamma}$	0.006	0.196	0.190	0.039	0.949	−0.021	0.202	0.186	0.041	0.932
	Naive	$\hat{\beta}_{1N}$	−0.017	0.158	0.150	0.025	0.932	−0.094	0.158	0.147	0.034	0.865
		$\hat{\beta}_{2N}$	−0.009	0.152	0.141	0.023	0.930	−0.049	0.153	0.140	0.026	0.907
		$\hat{\gamma}_N$	−0.234	0.178	0.159	0.087	0.649	−0.812	0.122	0.108	0.674	0.000
	Oracle	$\hat{\beta}_{1O}$	0.015	0.154	0.151	0.024	0.935	0.015	0.154	0.151	0.024	0.935
		$\hat{\beta}_{2O}$	0.007	0.150	0.142	0.022	0.939	0.007	0.150	0.142	0.022	0.939
		$\hat{\gamma}_O$	0.021	0.190	0.190	0.036	0.948	0.021	0.190	0.190	0.036	0.948
40%	Proposed	$\hat{\beta}_1$	0.015	0.177	0.172	0.032	0.941	0.010	0.178	0.172	0.032	0.938
		$\hat{\beta}_2$	0.002	0.169	0.163	0.029	0.946	−0.001	0.170	0.163	0.029	0.945
		$\hat{\gamma}$	0.015	0.225	0.215	0.051	0.944	−0.013	0.228	0.211	0.052	0.928
	Naive	$\hat{\beta}_{1N}$	−0.007	0.180	0.171	0.032	0.929	−0.074	0.181	0.168	0.038	0.892
		$\hat{\beta}_{2N}$	−0.010	0.171	0.162	0.029	0.936	−0.046	0.172	0.161	0.032	0.921
		$\hat{\gamma}_N$	−0.235	0.199	0.179	0.095	0.698	−0.822	0.134	0.122	0.693	0.001
	Oracle	$\hat{\beta}_{1O}$	0.021	0.175	0.172	0.031	0.940	0.021	0.175	0.172	0.031	0.940
		$\hat{\beta}_{2O}$	0.005	0.168	0.163	0.028	0.945	0.005	0.168	0.163	0.028	0.945
		$\hat{\gamma}_O$	0.028	0.220	0.215	0.049	0.948	0.028	0.220	0.215	0.049	0.948
80%	Proposed	$\hat{\beta}_1$	0.022	0.294	0.293	0.087	0.957	0.020	0.294	0.293	0.087	0.958
		$\hat{\beta}_2$	0.014	0.285	0.283	0.081	0.952	0.012	0.285	0.282	0.082	0.948
		$\hat{\gamma}$	0.023	0.383	0.360	0.147	0.934	−0.010	0.383	0.353	0.147	0.914
	Naive	$\hat{\beta}_{1N}$	0.011	0.294	0.292	0.087	0.951	−0.028	0.293	0.289	0.087	0.943
		$\hat{\beta}_{2N}$	0.006	0.285	0.282	0.081	0.943	−0.018	0.285	0.280	0.082	0.941
		$\hat{\gamma}_N$	−0.247	0.327	0.298	0.168	0.816	−0.850	0.223	0.202	0.773	0.035
	Oracle	$\hat{\beta}_{1O}$	0.026	0.293	0.293	0.086	0.955	0.026	0.293	0.293	0.086	0.955
		$\hat{\beta}_{2O}$	0.017	0.284	0.282	0.081	0.951	0.017	0.284	0.282	0.081	0.951
		$\hat{\gamma}_O$	0.036	0.376	0.360	0.143	0.941	0.036	0.376	0.360	0.143	0.941

The true value of the parameters $\beta_1 = 1$, $\beta_2 = 0.5$, $\gamma = 1.5$; $\phi(\cdot)$, the distortion function; CR, the censoring rate; Bias, the estimate value minus the true value; SD, the standard deviation; SE, the estimate of SD; MSE, the mean-square error; CP, empirical coverage percentage of the 95% confidence interval.

summarized in Tables 1 and 2 of the Supporting Information, from which we can see the proposed method also performs well even if the contamination structure is misspecified and performs better than the naive method.

5 | ANALYSIS WITH WILMS' TUMOR STUDY

We apply the proposed covariate-calibration method to the Wilms' tumor data, which were collected in two randomized studies in Wilms' tumor patients. Wilms' tumor is a rare kidney cancer occurring in young children. The National Wilms' Tumor Study Group (NWTSG) conducted several randomized studies to test different treatments in Wilms' tumor patients. We use a Wilms' tumor data including 3,915 patients participating in two of the NWTSG trials NWTS-3 and NWTS-4 (D'Angio et al., 1989; Green et al., 1998) to evaluate the joint effect of tumor weight, histological type, and other risk factors. The primary endpoint of the study was the survival time (in years). During the follow-up, 444 patients died of Wilms' tumor and the other 3,471 patients were censored, which led to the censoring rate of 88.66%. The mean observed time was 10.33 years (ranging from 0.01 to 22.50 years). We divide the data into two groups according to the histological type (favorable and unfavorable) and summarize the size and mean of each covariate in Table 5. It can be seen that 3,476 patients have favorable tumor, and the other 439 patients have unfavorable tumor. The mean observed time for patients with favorable tumor is 10.68 years, which is larger than the corresponding value (7.55) of the unfavorable tumor group.

TABLE 3 Simulation results for $\boldsymbol{\beta}$ and γ under sample size $n = 100$ and informative censoring mechanism

			$\phi(u) = (u+3)/7$					$\phi(u) = 3(u+1)^2/79$				
CR	**Method**	**Para.**	**Bias**	**SD**	**SE**	**MSE**	**CP**	**Bias**	**SD**	**SE**	**MSE**	**CP**
20%	Proposed	$\hat{\beta}_1$	0.021	0.221	0.216	0.049	0.945	0.007	0.222	0.215	0.049	0.946
		$\hat{\beta}_2$	0.007	0.206	0.199	0.043	0.936	0.001	0.208	0.199	0.043	0.940
		$\hat{\gamma}$	0.026	0.292	0.282	0.086	0.937	−0.041	0.296	0.271	0.089	0.899
	Naive	$\hat{\beta}_{1N}$	−0.006	0.222	0.215	0.049	0.938	−0.090	0.220	0.211	0.057	0.897
		$\hat{\beta}_{2N}$	−0.006	0.207	0.199	0.043	0.943	−0.044	0.207	0.198	0.045	0.932
		$\hat{\gamma}_N$	−0.211	0.261	0.239	0.112	0.794	−0.806	0.176	0.163	0.681	0.013
	Oracle	$\hat{\beta}_{1O}$	0.031	0.220	0.217	0.049	0.956	0.031	0.220	0.217	0.049	0.956
		$\hat{\beta}_{2O}$	0.011	0.206	0.199	0.042	0.934	0.011	0.206	0.199	0.042	0.934
		$\hat{\gamma}_O$	0.049	0.283	0.282	0.083	0.948	0.049	0.283	0.282	0.083	0.948
40%	Proposed	$\hat{\beta}_1$	0.034	0.244	0.239	0.061	0.944	0.021	0.245	0.238	0.060	0.943
		$\hat{\beta}_2$	0.010	0.232	0.221	0.054	0.938	0.004	0.233	0.220	0.054	0.933
		$\hat{\gamma}$	0.035	0.341	0.323	0.118	0.939	−0.036	0.342	0.311	0.119	0.905
	Naive	$\hat{\beta}_{1N}$	0.010	0.244	0.237	0.060	0.938	−0.067	0.242	0.233	0.063	0.917
		$\hat{\beta}_{2N}$	−0.003	0.231	0.220	0.053	0.940	−0.037	0.232	0.219	0.055	0.934
		$\hat{\gamma}_N$	−0.214	0.302	0.273	0.137	0.810	−0.819	0.199	0.186	0.711	0.030
	Oracle	$\hat{\beta}_{1O}$	0.044	0.243	0.240	0.061	0.946	0.044	0.243	0.240	0.061	0.946
		$\hat{\beta}_{2O}$	0.013	0.231	0.221	0.054	0.937	0.013	0.231	0.221	0.054	0.937
		$\hat{\gamma}_O$	0.053	0.331	0.323	0.113	0.948	0.053	0.331	0.323	0.113	0.948
80%	Proposed	$\hat{\beta}_1$	0.072	0.408	0.396	0.171	0.940	0.064	0.410	0.395	0.172	0.939
		$\hat{\beta}_2$	0.048	0.399	0.368	0.162	0.927	0.044	0.398	0.367	0.160	0.928
		$\hat{\gamma}$	0.103	0.589	0.571	0.357	0.938	0.021	0.588	0.548	0.346	0.934
	Naive	$\hat{\beta}_{1N}$	0.058	0.411	0.393	0.172	0.942	0.010	0.411	0.386	0.169	0.941
		$\hat{\beta}_{2N}$	0.038	0.393	0.367	0.156	0.935	0.015	0.395	0.365	0.157	0.930
		$\hat{\gamma}_N$	−0.176	0.530	0.478	0.312	0.888	−0.827	0.368	0.324	0.820	0.277
	Oracle	$\hat{\beta}_{1O}$	0.078	0.410	0.396	0.174	0.938	0.078	0.410	0.396	0.174	0.938
		$\hat{\beta}_{2O}$	0.050	0.402	0.369	0.164	0.931	0.050	0.402	0.369	0.164	0.931
		$\hat{\gamma}_O$	0.117	0.583	0.566	0.354	0.942	0.117	0.583	0.566	0.354	0.942

The true value of the parameters $\beta_1 = 1$, $\beta_2 = 0.5$, $\gamma = 1.5$; $\phi(\cdot)$, the distortion function; CR, the censoring rate; Bias, the estimate value minus the true value; SD, the standard deviation; SE, the estimate of SD; MSE, the mean-square error; CP, empirical coverage percentage of the 95% confidence interval.

Figure 1 shows the Kaplan–Meier curves for the two different tumor histological types, from which we can see that patients with favorable tumor experienced longer survival time.

The covariates included in this analysis are the weight of tumor-bearing specimen (abbreviated as wgt, in kilograms), the histological type of the tumor (type, being 0 if favorable and 1 otherwise), tumor stage (stage, coded by 1 and 0, indicating the spread of the tumor from localized to metastatic), age at diagnosis (age, measured in years), and the study number (num, 1 denotes NWTS-3 and 0 denotes NWTS-4).

We examine the following Cox proportional hazards regression model:

$$\lambda(t) = \lambda_0(t)\exp(\gamma \cdot \mathrm{wgt} + \beta_1 \cdot \mathrm{type} + \beta_2 \cdot \mathrm{stage} + \beta_3 \cdot \mathrm{age} + \beta_4 \cdot \mathrm{num}).$$

It is known that patients' wgt may be affected by their diam (i.e., the diameter of tumor); the scatter points of wgt versus diam shown in Figure 2 demonstrate that there exists a strong positive correlation between them. Here we directly adjust the potential distorting covariate with the proposed method and assume that the distortion model as $\widetilde{\mathrm{wgt}} = \phi(\mathrm{diam}) \cdot \mathrm{wgt}$, where $\phi(\cdot)$ is an unknown link function and $\widetilde{\mathrm{wgt}}$ is the observed wgt. The analysis results of the covariate effects are summarized in Table 6. As a comparison, we also present the results of the naive method which ignore the contamination of wgt. By observing the results, the p-value of wgt is 0.008 for our proposed method, which means wgt has significant influence on patients' survival time, while the corresponding value is 0.244 for the naive method. From the medical standpoint,

TABLE 4 Simulation results for $\boldsymbol{\beta}$ and γ under sample size $n = 200$ and informative censoring mechanism

			$\phi(u) = (u+3)/7$					$\phi(u) = 3(u+1)^2/79$				
CR	**Method**	**Para.**	**Bias**	**SD**	**SE**	**MSE**	**CP**	**Bias**	**SD**	**SE**	**MSE**	**CP**
20%	Proposed	$\widehat{\beta}_1$	0.009	0.150	0.146	0.023	0.938	0.003	0.151	0.145	0.023	0.937
		$\widehat{\beta}_2$	0.006	0.145	0.136	0.021	0.942	0.003	0.146	0.136	0.021	0.939
		$\widehat{\gamma}$	0.011	0.199	0.191	0.040	0.946	−0.016	0.204	0.188	0.042	0.936
	Naive	$\widehat{\beta}_{1N}$	−0.018	0.152	0.144	0.023	0.935	−0.097	0.151	0.142	0.032	0.861
		$\widehat{\beta}_{2N}$	−0.007	0.147	0.135	0.022	0.932	−0.047	0.148	0.135	0.024	0.912
		$\widehat{\gamma}_N$	−0.229	0.181	0.160	0.085	0.663	−0.807	0.124	0.109	0.667	0.000
	Oracle	$\widehat{\beta}_{1O}$	0.015	0.149	0.146	0.022	0.938	0.015	0.149	0.146	0.022	0.938
		$\widehat{\beta}_{2O}$	0.008	0.145	0.136	0.021	0.940	0.008	0.145	0.136	0.021	0.940
		$\widehat{\gamma}_O$	0.026	0.193	0.192	0.038	0.949	0.026	0.193	0.192	0.038	0.949
40%	Proposed	$\widehat{\beta}_1$	0.010	0.165	0.161	0.027	0.939	0.005	0.166	0.161	0.028	0.940
		$\widehat{\beta}_2$	0.007	0.159	0.150	0.025	0.948	0.004	0.160	0.150	0.026	0.940
		$\widehat{\gamma}$	0.015	0.221	0.218	0.049	0.948	−0.013	0.227	0.214	0.052	0.932
	Naive	$\widehat{\beta}_{1N}$	−0.012	0.167	0.160	0.028	0.940	−0.082	0.168	0.157	0.035	0.889
		$\widehat{\beta}_{2N}$	−0.005	0.161	0.150	0.026	0.934	−0.040	0.163	0.149	0.028	0.918
		$\widehat{\gamma}_N$	−0.231	0.201	0.183	0.094	0.696	−0.816	0.140	0.124	0.685	0.000
	Oracle	$\widehat{\beta}_{1O}$	0.017	0.162	0.161	0.027	0.943	0.017	0.162	0.161	0.027	0.943
		$\widehat{\beta}_{2O}$	0.009	0.159	0.150	0.025	0.944	0.009	0.159	0.150	0.025	0.944
		$\widehat{\gamma}_O$	0.030	0.217	0.219	0.048	0.950	0.030	0.217	0.219	0.048	0.950
80%	Proposed	$\widehat{\beta}_1$	0.030	0.259	0.256	0.068	0.942	0.026	0.259	0.256	0.068	0.939
		$\widehat{\beta}_2$	0.020	0.254	0.243	0.065	0.944	0.019	0.255	0.243	0.065	0.945
		$\widehat{\gamma}$	0.057	0.397	0.376	0.161	0.936	0.023	0.392	0.368	0.154	0.931
	Naive	$\widehat{\beta}_{1N}$	0.015	0.261	0.255	0.068	0.939	−0.029	0.262	0.252	0.069	0.928
		$\widehat{\beta}_{2N}$	0.012	0.257	0.243	0.066	0.945	−0.013	0.261	0.241	0.068	0.937
		$\widehat{\gamma}_N$	−0.219	0.343	0.312	0.165	0.841	−0.831	0.235	0.211	0.746	0.062
	Oracle	$\widehat{\beta}_{1O}$	0.034	0.257	0.256	0.067	0.942	0.034	0.257	0.256	0.067	0.942
		$\widehat{\beta}_{2O}$	0.022	0.254	0.244	0.065	0.944	0.022	0.254	0.244	0.065	0.944
		$\widehat{\gamma}_O$	0.069	0.400	0.375	0.165	0.931	0.069	0.400	0.375	0.165	0.931

The true value of the parameters $\beta_1 = 1$, $\beta_2 = 0.5$, $\gamma = 1.5$; $\phi(\cdot)$, the distortion function; CR, the censoring rate; Bias, the estimate value minus the true value; SD, the standard deviation; SE, the estimate of SD; MSE, the mean-square error; CP, empirical coverage percentage of the 95% confidence interval.

TABLE 5 The data of the NWTSG trials grouped by the histological type

	Overall	**Favorable**	**Unfavorable**
size	3915	3476	439
wgt	604.56	603.74	611.12
diam	11.21	11.20	11.32
age	3.53	3.52	3.68
stage (%)	64.78	66.28	52.85
num (%)	42.68	42.55	43.74
time	10.33	10.68	7.55
cen.rate (%)	88.66	92.23	60.36

Overall, the total patients; favorable, the patients with favorable tumor; unfavorable, the patients with unfavorable tumor; size, the sample size; wgt, the mean weight of tumor-bearing specimens; diam, the mean diameter of tumors; age, the mean age of patients at diagnosis; stage, the percentage of patients with tumor localized spread; num, the percentage of patients in NWTS-3 trial; time, the mean observed time; cen.rate, the censoring rate.

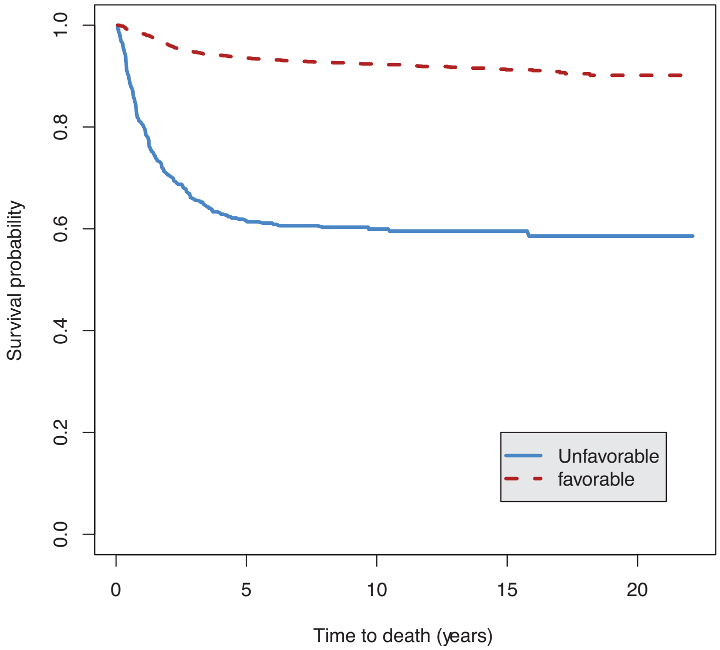

FIGURE 1 Kaplan–Meier survival curves stratified by two different histological types of the tumor in in the NWTSG trials

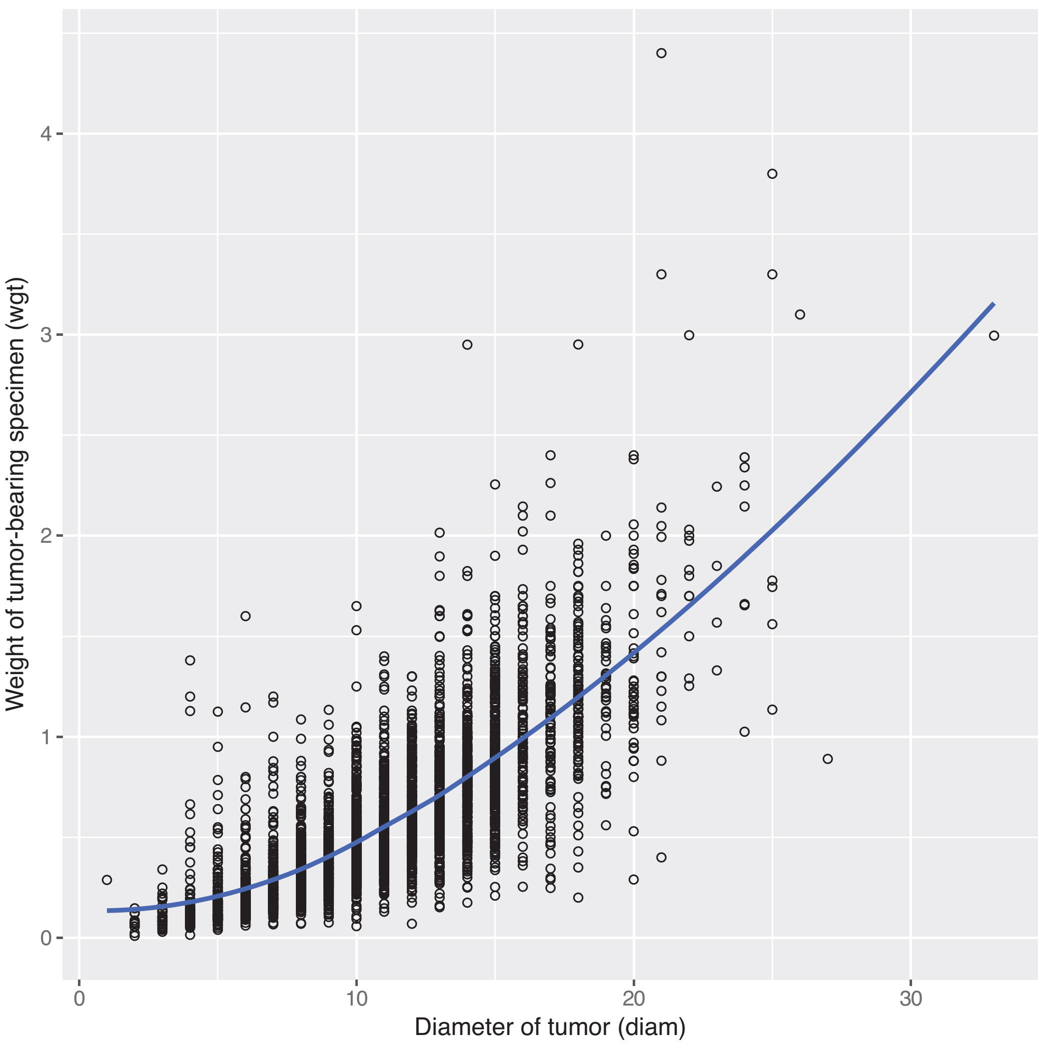

FIGURE 2 The scatter diagram of tumor-bearing specimen's weight (wgt) versus tumor's diameter (diam) for the NWTSG trials

TABLE 6 The analysis results of the covariate effects in the NWTSG trials under the cases which includes and not includes "diag" as a covariate in the model

	Method	Covariate	EST	SE	*P*-value
Case 1	Proposed	wgt	−0.482	0.180	0.008
		type	1.820	0.096	<0.001
		stage	−0.900	0.097	<0.001
		age	0.070	0.020	<0.001
		num	0.171	0.098	0.081
	Naive	wgt	−0.139	0.119	0.244
		type	1.821	0.096	<0.001
		stage	−0.908	0.099	<0.001
		age	0.066	0.020	0.001
		num	0.187	0.097	0.055
Case 2	Proposed	wgt	−0.485	0.190	0.011
		type	1.820	0.101	<0.001
		stage	−0.907	0.050	<0.001
		age	0.072	0.021	<0.001
		num	0.174	0.097	0.074
		diam	−0.005	0.014	0.726
	Naive	wgt	−0.253	0.170	0.136
		type	1.822	0.101	<0.001
		stage	−0.898	0.049	<0.001
		age	0.068	0.021	0.001
		num	0.180	0.097	0.064
		diam	0.017	0.018	0.364

Case 1: does not include "diam" as a covariate in the model; Case 2: include "diam" as a covariate in the model; wgt: the weight of tumor-bearing specimen; type: the histological type of the tumor; stage: the tumor stage; age: the age of patients at diagnosis; num: the study number; EST: the estimate of the parameters; SE: the standard error estimate; *P*-value: the *p*-value of the parameters.

wgt has great influence on patients' survival time, whereas ignoring the contamination leads to this covariate insignificant. Furthermore, from all these two methods, we can conclude that patients with favorable tumor will possess longer survival time compared with ones with unfavorable tumor, which coincides with Figure 1.

For comparison, we further analyze these data by including diam as a covariate in the model. The corresponding results of the proposed method and the naive method are also summarized in Table 6, from which we can draw similar conclusions, that is, the naive method which ignores the contamination leads to wgt insignificant. As we all know, both wgt and diam have great influence on patients' survival time; however, the results show that if we add these two variables to the model simultaneously then it will lead to diam insignificant. This phenomenon may be caused by the strong correlation between diam and wgt, so we do not recommend adding the confounding variable "diam" to the model.

6 | CONCLUSION

The covariate-adjusted problem is a common contamination problem in biomedical studies. Similar issues are also reported in other fields, for example, in environmental studies exposures are often calibrated by the daily environment or ambient measures, like the role of BMI in medical studies, or genomic studies where the library size is being normalized. Our method deals with the type of some primary covariates that are observed after being distorted by a multiplicative factor (an unknown function of an observable confounding variable). We fill in the gap in the literature on censored survival data with the distorting function in a primary risk factor, which is lacking in terms of the statistical method. We propose a direct estimation procedure to estimate the regression parameters in the Cox proportional hazards regression model. Numerical results show that the proposed method is working very well in correcting the bias arising from covariate distortion. It performs stably to a variety of distortion functions. An important improvement of our method is that we allow flexible distorting models to handle various confounding mechanisms. It is easy to compute and will provide a critical tool for researchers facing with this type data in practice. The proposed method is actually a two-step procedure, we first obtain a consistent estimator of the distorted covariate by employing the kernel smoothing method and then obtain the parameter estimation by plugging in the estimated covariate. In the first step, we can also employ other methods to estimate the distorted covariate, such as spline or local polynomial approximation.

A few remarks on using the proposed method in real studies. First, on the construction of confidence interval of the proposed estimation, we note that because of the nonlinear structure of the estimated partial likelihood and the maximum partial likelihood estimation do not have a closed form, the establishment of theoretical properties in this paper is more difficult than the linear model. The asymptotic covariance matrix derived in Theorem 3.1 depends on several unknown components; therefore, it is difficult to construct confidence region based on normal approximation. We recommend to use the common sandwich approach to obtain the standard error estimation, which has been tested and demonstrated to perform well in our numerical studies.

Second, for ease of exposition, we consider only one confounding variable. In many applications, there may exist multiple distortion variables simultaneously affecting the primary covariate. The proposed method can handle this case, and the sandwich method can also be employed to obtain the standard error estimation. But deriving theoretical properties of the corresponding estimators will be more difficult and need additional technicalities.

Finally, if we require to divide the distorted variable by the estimated distorting function, we impose some regularity assumptions on the curve of the distorting function. In particular, the proposed method cannot be applied if $E(X)$ vanishes. Delaigle et al. (2016) proposed a more flexible nonparametric estimator for the regression function, which significantly weakens some of the strong assumptions on the distorting function. Further research is underway to extend this work to censored survival data.

ACKNOWLEDGMENTS

This work was partially supported by the National Natural Science Foundation of China (no. 11771366, no. 11971362, no. 11671311, no. 11901581, no. 12071483) and U.S. National Institute of Health grant (P42ES031007, P30ES010126).

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in R package “addhazard”.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge “**Reproducible Research**” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

ORCID

Yanyan Liu https://orcid.org/0000-0003-2304-4877
Jing Zhang https://orcid.org/0000-0002-0368-7739

REFERENCES

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, *71*, 353–360.

Chen, Y.-C., Genovese, C. R., Tibshirani, R. J., & Wasserman, L. (2016). Nonparametric modal regression. *The Annals of Statistics*, *44*, 489–514.

Chen, Y.-C., Genovese, C. R., & Wasserman, L. (2018). Density level sets: Asymptotics, inference, and visualization. *Journal of the American Statistical Association*, *112*, 1684–1696.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, *34*, 187–220.

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

Cui, X., Guo, W. S., Lin, L., & Zhu, L. X. (2009). Covariate-adjusted nonlinear regression. *The Annals of Statistics*, *37*, 1839–1870.

D'angio, G. J., Breslow, N., Beckwith, J. B., Evans, A., Baum, E., Delorimier, A., Fernbach, D., Hrabovsky, E., Jones, B., Kelalis, P., Othersen, H. B., Tefft, M., Thomas, P. R. M. (1989). Treatment of Wilms' tumor: Results of the third national Wilms' tumor study. *Cancer*, *64*, 349–360.

Delaigle, A., Hall, P., & Zhou, W. X. (2016). Nonparametric covariate-adjusted regression. *The Annals of Statistics*, *44*, 2190–2220.

Fan, J., & Zhang, J. (2000). Two-step estimation of functional linear models with applications to longitudinal data. *Journal of the Royal Statistical Society, Series B*, *62*, 303–322.

Giné, E., & Guillou, A. (2002). Rates of strong uniform consistency for multivariate kernel density estimators. *Annales de l'Institut Henri Poincaré*, *38*, 907–921.

Green, D. M., Breslow, N. E., Beckwith, J. B., Finklestein, J. Z., Grundy, P. E., Thomas, P. R., Kim, T., Shochat, S. J., Haase, G. M., Ritchey, M. L., Kelalis, P. P., D'Angio, G. J. (1998). Comparison between single-dose and divided-dose administration of dactinomycin and doxorubicin for patients with Wilms' tumor: A report from the National Wilms' Tumor Study Group. *Journal of Clinical Oncologys*, *16*, 237–245.

Huang, Y. J., & Wang, C. Y. (2000). Cox regression with accurate covariates unascertainable: A nonparametric-correction approach. *Journal of the American Statistical Association*, *95*, 1209–1219.

Hu, C., & Lin, D. Y. (2002). Cox regression with covariate measurement error. *Scandinavian Journal of Statistics*, *29*, 637–655.

Hwang, J. T. (1986). Multiplicative errors-in-variables models with applications to recent data released by the U.S. Department of Energy. *Journal of the American Statistical Association*, *81*, 680–688.

Kaysen, G. A., Dubin, J. A., Müller, H. G., Mitch, W. E., Rosales, L. M., Levin, N. W., & Hemo Study Group. (2002). Relationships among inflammation nutrition and physiologic mechanisms establishing albumin levels in hemodialysis patients. *Kidney International*, *61*, 2240–2249.

Li, Y., & Ryan, L. (2004). Survival analysis with heterogeneous covariate measurement error. *Journal of the American Statistical Association*, *99*, 724–735.

Nguyen, D. V., & Sentürk, D. (2008). Multicovariate-adjusted regression models. *Journal of Statistical Computation and Simulation*, *78*, 813–827.

Prentice, R. L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, *69*, 331–342.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, *9*, 65–78.

Sentürk, D., & Müller, H. G. (2005). Covariate-adjusted regression. *Biometrika*, *92*, 75–89.

Sentürk, D., & Müller, H. G. (2006). Inference for covariate-adjusted regression via varying coefficient models. *The Annals of Statistics*, *34*, 654–679.

Sentürk, D., & Müller, H. G. (2009). Covariate-adjusted generalized linear models. *Biometrika*, *96*, 357–370.

Wang, C. Y., Hsu, L., Feng, Z. D., & Prentice, R. L. (1997). Regression calibration in failure time regression. *Biometrics*, *53*, 131–145.

Wu, C. F. (1981). Asymptotic theory of nonlinear least squares estimation. *The Annals of Statistics*, *9*, 501–513.

Zhang, J., Yu, Y., Zhu, L. X., & Liang, H. (2013). Partial linear single index models with distortion measurement errors. *Annals of the Institute of Statistical Mathematics*, *65*, 237–267.

Zhang, J., Zhu, L. X., & Liang, H. (2012). Nonlinear models with measurement errors subject to single-indexed distortion. *Journal of Multivariate Analysis*, *112*, 1–23.

Zhang, J., Zhu, L. P., & Zhu, L. X. (2012). On a dimension reduction regression with covariate adjustment. *Journal of Multivariate Analysis*, *104*, 39–55.

Zhou, H., & Pepe, M. S. (1995). Auxiliary covariate data in failure time regression. *Biometrika*, *82*, 139–149.

Zhou, H., & Wang, C. Y. (2000). Failure time regression with continuous covariates measured with error. *Journal of the Royal Statistical Society, Series B*, *62*, 657–665.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Liu Y, Wu Y, Zhang J, Zhou H. Cox regression analysis for distorted covariates with an unknown distortion function. *Biometrical Journal*. 2021;1–16. https://doi.org/10.1002/bimj.202000209

APPENDIX A: REGULARITY CONDITIONS

Unless otherwise stated, all limits are taken as $n \to \infty$. Suppose $\mathbf{a} = (a_1, \dots, a_p)^{\mathrm{T}}$ and $\mathbf{b} = (b_1, \dots, b_p)^{\mathrm{T}}$ are p-vectors, then we write $\mathbf{a} \otimes \mathbf{b}$ for the matrix $\mathbf{ab}^{\mathrm{T}}$. Also we write $\mathbf{a}^{\otimes 2}$ for the matrix $\mathbf{a} \otimes \mathbf{a}$. For a matrix $\mathbf{A}$ or vector $\mathbf{a}$, let $\|\mathbf{A}\| = \sup_{i,j} |a_{ij}|$ and $\|\mathbf{a}\| = \sup_i |a_i|$. For matrix or vector sequences $\mathbf{A}_n$ and $\mathbf{B}_n$, denote $\mathbf{A}_n \xrightarrow{p} \mathbf{A}$ if $\|\mathbf{A}_n - \mathbf{A}\| \xrightarrow{p} 0$ and denote $A_n = B_n + o_p(1)$ if $\|A_n - B_n\| \xrightarrow{p} 0$. Denote $|\mathbf{a}| = (\sum a_i^2)^{1/2}$ and $\mathrm{diag}(\mathbf{a})$ as the diagonal matrix whose diagonal vector is $\mathbf{a}$. We set $\theta = (\boldsymbol{\beta}^{\mathrm{T}}, \gamma)^{\mathrm{T}}$, $\mathbf{V} = (\mathbf{Z}^{\mathrm{T}}, X)^{\mathrm{T}}$, $N_i(t) = I(\widetilde{T}_i \le t, \Delta_i = 1)$, $\bar{N} = \sum_{i=1}^n N_i$, and $Y_i(t) = I(\widetilde{T}_i \ge t)$. Let τ denote the end time of the study. Here, we introduce the following notations:

$$S^{(l)}(\theta, t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{V}_i^{\otimes l} Y_i(t) \exp\left(\mathbf{V}_i^{\mathrm{T}}\theta\right),$$

$$E(\theta, t) = \frac{S^{(1)}(\theta, t)}{S^{(0)}(\theta, t)},$$

$$V(\theta, t) = \frac{S^{(2)}(\theta, t)}{S^{(0)}(\theta, t)} - E(\theta, t)^{\otimes 2},$$

for $l = 0, 1, 2$. Note that $S^{(0)}(\theta, t)$ is a scalar, $S^{(1)}(\theta, t)$ and $E(\theta, t)$ are $(p+1)$-vectors, and $S^{(2)}(\theta, t)$ and $V(\theta, t)$ are $(p+1) \times (p+1)$ matrices. Before proving the theorem, we first describe the regular conditions needed as follows:

C1. (Finite interval). $\int_0^\tau \lambda_0(t)\mathrm{d}t < \infty$.

C2. (Asymptotic stability). There exists a neighborhood $\mathscr{B}$ of θ_0, and scalar, vector, and matrix functions $s^{(0)}$, $s^{(1)}$ and $s^{(2)}$ defined on $\mathscr{B} \times [0, \tau]$ such that for $j = 0, 1, 2$,

$$\sup_{t\in[0,\tau],\theta\in\mathscr{B}} \|S^{(j)}(\theta, t) - s^{(j)}(\theta, t)\| \xrightarrow{p} 0.$$

C3. (Lindeberg condition). There exists $\delta > 0$ such that

$$n^{-1/2}\sup_{i,t} |\mathbf{V}_i|\, Y_i(t) I\left\{\theta_0^{\mathrm{T}}\mathbf{V}_i > -\delta\,|\mathbf{V}_i|\right\} \xrightarrow{p} 0.$$

C4. (Asymptotic regularity conditions). Let $\mathscr{B}$, $s^{(0)}$, $s^{(1)}$, and $s^{(2)}$ be as in condition C2 and define $e = s^{(1)}/s^{(0)}$ and $v = s^{(2)}/s^{(0)} - e^{\otimes 2}$. For all $\theta \in \mathscr{B}$, $t \in [0, \tau]$:

$$s^{(1)}(\theta, t) = \frac{\partial}{\partial\theta}s^{(0)}(\theta, t),\ s^{(2)}(\theta, t) = \frac{\partial^2}{\partial\theta^2}s^{(0)}(\theta, t),$$

$s^{(0)}(\cdot, t)$, $s^{(1)}(\cdot, t)$, and $s^{(2)}(\cdot, t)$ are continuous functions of $\theta \in \mathscr{B}$, uniformly in $t \in [0, \tau]$, $s^{(0)}$, $s^{(1)}$, and $s^{(2)}$ are bounded on $\mathscr{B} \times [0, \tau]$, $s^{(0)}$ is bounded away from zero on $\mathscr{B} \times [0, \tau]$, and the matrix

$$\Sigma = \int_0^\tau v(\theta_0, t)s^{(0)}(\theta_0, t)\lambda_0(t)\,\mathrm{d}t$$

is positive definite.

C5. $p(u)$ and $\phi(u)$ are bounded away from zero and have bounded second derivatives.

C6. $\int_{-\infty}^{\infty} K(x)\,\mathrm{d}x = 1$, $\int_{-\infty}^{\infty} xK(x)\,\mathrm{d}x = 0$ and $\int_{-\infty}^{\infty} x^2K(x)\,\mathrm{d}x < \infty$.

C7. The kernel function satisfies condition K_1 in Giné and Guillou (2002). Let

$$\mathscr{K} = \left\{y \mapsto K(\frac{x-y}{h_n}) : x \in R, h_n > 0\right\},$$

then for any $\epsilon > 0$, $\mathscr{K}$ satisfies that

$$\sup_P N\left(\mathscr{K}, L_2(P), \epsilon\|F\|_{L_2(P)}\right) \leq \left(\frac{A}{\epsilon}\right)^\nu$$

for some positive constants A and ν, where $N(\Omega, d, \epsilon)$ denotes the ϵ-covering number of the metric space (Ω, d), F is the envelope function of $\mathscr{K}$, the supremum is taken over R, and the norm $\|F\|^2_{L_2(P)}$ is defined as $\int_R |F(x)|^2\mathrm{d}P(x)$.

C8. $|\log h_n|/\log\log n \to \infty$ and $nh_n/|\log h_n| \to \infty$, h_n and $(nh_n)^{-1}$ monotonically converge to zero as $n \to \infty$.

C9. $E(X)$ and $E(Z_i)$ $(i = 1, \ldots, p)$ are bounded away from 0.

These conditions are mild and can be satisfied in most of circumstances. Conditions C1–C4 are essential for the asymptotic results of the Cox proportional hazards regression model. Condition C5 is a mild smoothness condition on the involved functions. Condition C6 is common for a kernel function, and C7 is to regularize the complexity of the kernel function so that the supremum norm for kernel functions can be bounded in probability, which are also imposed in Chen et al. (2016, 2018). Specially, the Gauss kernel function satisfies the conditions C6 and C7. Condition C8 states that the bandwidth h_n converges to zero at a certain rate with respect to the sample size n. Condition C9 is necessary in the study of covariate-adjusted problems; see Sentürk and Müller (2006).

APPENDIX B: PROOFS OF ASYMPTOTIC PROPERTIES

As a preparation, we state a lemma, which is extracted from Lemma B.2 of Zhang et al. (2012) and frequently used in the process of the proof.

Lemma B.1. *Let $\eta(\mathbf{z})$ be a continuous function satisfying $E[\eta(\mathbf{Z})]^2 < \infty$. Assume that conditions C5–C9 hold. The following asymptotic representation holds:*

$$\frac{1}{n}\sum_{i=1}^{n}(\widehat{X}_i - X_i)\eta(\mathbf{Z}_i) = \frac{1}{n}\sum_{i=1}^{n}(\widetilde{X}_i - X_i)\frac{E[X\eta(\mathbf{Z})]}{E(X)} + o_p(n^{-1/2}).$$

Proof of Theorem 3.1

Proof of (i). Denoted by $\theta = (\boldsymbol{\beta}^{\mathrm{T}}, \gamma)^{\mathrm{T}}$, $\mathbf{V} = (\mathbf{Z}^{\mathrm{T}}, X)^{\mathrm{T}}$, and $\widehat{\mathbf{V}} = (\mathbf{Z}^{\mathrm{T}}, \widehat{X})^{\mathrm{T}}$, the log partial likelihood of this covariate-adjusted Cox model can be written as

$$\widehat{L}_n(\boldsymbol{\beta}, \gamma) = \sum_{i=1}^{n}\int_0^{\tau}\widehat{\mathbf{V}}_i^{\mathrm{T}}\theta \, \mathrm{d}N_i(t) - \int_0^{\tau}\log\left\{\sum_{i=1}^{n}Y_i(t)\exp(\widehat{\mathbf{V}}_i^{\mathrm{T}}\theta)\right\}\mathrm{d}\bar{N}(t).$$

Set

$$L_n(\boldsymbol{\beta}, \gamma) = \sum_{i=1}^{n}\int_0^{\tau}\mathbf{V}_i^{\mathrm{T}}\theta \, \mathrm{d}N_i(t) - \int_0^{\tau}\log\left\{\sum_{i=1}^{n}Y_i(t)\exp(\mathbf{V}_i^{\mathrm{T}}\theta)\right\}\mathrm{d}\bar{N}(t).$$

The main point of the proof lies in stating that, for any $\theta \in \Theta$,

$$\widehat{L}_n(\boldsymbol{\beta}, \gamma) - L_n(\boldsymbol{\beta}, \gamma) = o_p(n).$$

This implies, by the fact, that $\widehat{\theta} = \operatorname{argmax}_{\theta\in\Theta}\widehat{L}_n(\boldsymbol{\beta}, \gamma)$ and the consistency of the Cox model under conditions C1–C4, and the consistency of $\widehat{\theta}$ follows from Lemma 1 of Wu (1981). The detailed proof is given in the Supporting Information. □

Proof of (ii). Let

$$\widehat{U}(\theta) = \sum_{i=1}^{n}\int_0^{\tau}\widehat{\mathbf{V}}_i \, \mathrm{d}N_i(t) - \int_0^{\tau}\frac{\sum_{i=1}^{n}Y_i(t)\widehat{\mathbf{V}}_i \cdot \exp(\widehat{\mathbf{V}}_i^{\mathrm{T}}\theta)}{\sum_{i=1}^{n}Y_i(t)\exp(\widehat{\mathbf{V}}_i^{\mathrm{T}}\theta)}\mathrm{d}\bar{N}(t).$$

By the Taylor expansion, there exists θ^* between θ_0 and $\widehat{\theta}$ such that

$$\frac{1}{\sqrt{n}}\widehat{U}(\widehat{\theta}) - \frac{1}{\sqrt{n}}\widehat{U}(\theta_0) = \frac{1}{n}\frac{\partial\widehat{U}(\theta^*)}{\partial\theta}\sqrt{n}(\widehat{\theta} - \theta_0).$$

By the definition of $\widehat{\theta}$, we know that $\widehat{U}(\widehat{\theta}) = \mathbf{0}$. So we have

$$\sqrt{n}(\widehat{\theta} - \theta_0) = \left\{ -\frac{1}{n}\frac{\partial \widehat{U}(\theta^*)}{\partial \theta} \right\}^{-1} \cdot \frac{1}{\sqrt{n}}\widehat{U}(\theta_0).$$

We can prove that

$$-\frac{1}{n}\frac{\partial \widehat{U}(\theta^*)}{\partial \theta} \xrightarrow{p} \Sigma, \tag{B.1}$$

and

$$\frac{1}{\sqrt{n}}\widehat{U}(\theta_0) \xrightarrow{d} \mathrm{N}(0, \Sigma + \Omega), \tag{B.2}$$

where $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$ is defined in condition C4, $\zeta = (-\Sigma_{12}^{\mathrm{T}}\gamma_0, -\Sigma_{22}\gamma_0)^{\mathrm{T}}$, and $\Omega = \frac{\mathrm{Var}(\widetilde{X}) - \mathrm{Var}(X)}{\{E(X)\}^2}\zeta\zeta^{\mathrm{T}}$.

Combining (B.1) and (B.2), we have

$$\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{d} \mathrm{N}(0, \Sigma^{-1}(\Sigma + \Omega)\Sigma^{-1}),$$

where

$$\begin{aligned}
\Sigma^{-1}(\Sigma + \Omega)\Sigma^{-1} &= \Sigma^{-1}\Sigma\Sigma^{-1} + \frac{\mathrm{Var}(\widetilde{X}) - \mathrm{Var}(X)}{\{E(X)\}^2}\Sigma^{-1}\zeta\zeta^{\mathrm{T}}\Sigma^{-1} \\
&= \Sigma^{-1}\Sigma\Sigma^{-1} + \frac{\mathrm{Var}(\widetilde{X}) - \mathrm{Var}(X)}{\{E(X)\}^2}\begin{pmatrix} 0 & \cdots & 0 & 0 \\ \vdots & & \vdots & 0 \\ 0 & \cdots & 0 & 0 \\ 0 & \cdots & 0 & \gamma_0^2 \end{pmatrix}.
\end{aligned}$$

We can obtain that

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathrm{N}\big(\mathbf{0}, (\Sigma^{-1})_p\big),$$

and

$$\sqrt{n}(\widehat{\gamma} - \gamma_0) \xrightarrow{d} \mathrm{N}\left(0, (\Sigma^{-1})_{(p+1,p+1)} + \frac{\mathrm{Var}(\widetilde{X}) - \mathrm{Var}(X)}{\{E(X)\}^2}\gamma_0^2\right),$$

where $(\Sigma^{-1})_p$ and $(\Sigma^{-1})_{(p+1,p+1)}$, respectively, represent the pth-order sequential principal minor and the $(p+1)$th diagonal element of matrix Σ^{-1}. The detailed proof of (B.1) and (B.2) is given in the Supporting Information. □