Research Article

Regression Analysis for Outcome-Dependent Sampling Design under the Covariate-Adjusted Additive Hazards Model

Yingli Pan,¹ Songlin Liu ,¹ Yanli Zhou ,² and Guangyu Song¹

¹Hubei Key Laboratory of Applied Mathematics, Faculty of Mathematics and Statistics, Hubei University, Wuhan 430062, China

²School of Finance, Zhongnan University of Economics and Law, Wuhan 430073, China

Correspondence should be addressed to Songlin Liu; liusonglin@hubu.edu.cn and Yanli Zhou; yanli.zhou1517@zuel.edu.cn

Received 26 July 2020; Revised 2 September 2020; Accepted 13 September 2020; Published 30 October 2020

Academic Editor: Tongqian Zhang

Copyright © 2020 Yingli Pan et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This paper provides a new insight into an economical and effective sampling design method relying on the outcome-dependent sampling (ODS) design in large-scale cohort research. Firstly, the importance and originality of this paper is that it explores how to fit the covariate-adjusted additive Hazard model under the ODS design; secondly, this paper focused on estimating the distortion function through nonparametric regression and required observation of the covariate on the confounding factors of distortion; moreover, this paper further calibrated the contaminated covariates and proposed the estimators of the parameters by analyzing the calibrated covariates; finally, this paper established the large sample property and asymptotic normality of the proposed estimators and conducted many more simulations to evaluate the finite sample performance of the proposed method. Empirical research demonstrates that the results from both artificial and real data verified good performance and practicality of the proposed ODS method in this paper.

1. Introduction

Generally, the major cost of studying large cohort is tied up in collecting the expensive exposure variables, which casts a poor shadow over the researchers with the burden of limited budgets. Therefore, the simple random sampling design with the characteristics of cost expensive and time consuming has been risking losing ground in the long run. To achieve some certain goal, it is no wonder that many cost-effective strategies have been invoked. In the early 1980s, Prentice [1] proposed the notation of case-cohort design, and the exposure variables were measured on a simple random sample, which is called a subcohort, as well as all the cases that experienced events that we were interested in. Since then, the application of case-cohort sampling in the survival analysis has been reported by Self and Prentice [2], Tsai [3], and Kim et al. [4]. The case-cohort design is expected to be economical and effective sampling techniques for rare events. When the censoring rate is relatively low or medium, the method of generalized case-cohort design has been

developed in responding to lowering the research cost. In addition to randomly selecting a subcohort from the entire cohort, the information containing the relevant covariables is collected for only a subset of the failure individuals (e.g., Chen [5], Cai and Zeng [6], and Kang and Cai [7]).

As a matter of fact, the failure time outcome-dependent sampling (ODS) design is considered to be another economical and effective alternative to the simple random sampling design. Exposure variables are measured against samples from two components, the subcohort and additional supplementary samples (see Chatterjee et al. [8], Zhou et al. [9], and Weaver and Zhou [10]). Regarding the ODS data, massive research has been carried out and formed a wealth of literatures. For completely observed data, the studies by Zhou et al. [11, 12] and Qin and Zhou [13] offer the comprehensive analysis of the inference methods based on the partially regression linear models for data from the ODS design. The key systematic study of how to fit the generalized linear model with the data obtained from two-stage ODS design was reported by Yan et al. [14]. For

censored data, detail examination of the ODS design by Ding et al. [15] showed the estimated impact of environmental pollutants on women's subfertility. A significant discussion on estimating the equation method with an ODS sampling scheme under Cox's proportional hazards model was presented by Yu et al. [16].

In clinical trials and biomedical, instead of direct observation, the covariates are observed by multiplication of unknown functions of an observable confounder. As a result, the regression with contaminated covariates was originally derived by Şentürk and Müller [17, 18], in which the contamination information for model covariates cannot be ignored; otherwise, it will result in a biased estimator and the statistical inference may be misled. Since then, numerous extensions in various aspects have been developed by Şentürk and Müller [19], Cui et al. [20], and Li et al. [21]. For completely observed data, one study by Cui [22] proposed to use the nonparametric kernel estimation method to calibrate the contaminated variables and then conducted parameter estimation under the covariate-adjusted linear model, and further research conducted by Zhang et al. [23] extended the method to the nonlinear models with contaminated variables. A key study by Delaigle et al. [24] derived the process of several nonparametric covariate-adjusted estimators of conditional mean function. A preliminary nonparametric test for covariate-adjusted models was undertaken by Zhao and Xie [25], who found that the proposed test statistic has the same limit distribution as the response and predictors are supposed to be obtained directly. For survival data with censoring, very few studies have investigated survival models with contaminated covariates and even fewer people are ready to tackle the challenge in the ODS design we are discussing here.

We study the following covariate-adjusted additive hazards model under the ODS design in this paper:

$$\begin{cases} \lambda(t|Z) = \lambda_0(t) + \theta^T H + \gamma X, \\ W = \phi(U)X, \end{cases} \quad (1)$$

where $\lambda_0(t)$ is the unknown baseline hazard function, $\beta = (\theta^T, \gamma^T)^T$ is the unknown parameter of p -dimension, θ and γ are $(p-1)$ -dimensional and 1-dimensional parameter, respectively, $Z = (H^T, X)^T$ is the p -dimensional covariate, H is the observed $(p-1)$ -dimensional covariate, X is the unobservable 1-dimensional covariate, W is the actual observed 1-dimensional covariate, and $\phi(\cdot)$ is the unknown distorting function of observable confounding variable U . We focus on the method of nonparametric kernel estimation to obtain the estimator of the distortion function and calibrate the covariate X . Meanwhile, we attempt to weigh the contributions of the subcohort and the supplemental sample differently, which resulted in a weighted estimation equation with the help of calibrated covariates. Owing to the ODS design and the covariate-adjusted process, it happens to be challengeable in the work of theoretical developments. To overcome the challenge, it will be followed by an approximation to the weighted estimation equation, which is taken as the main basis for obtaining the theory properties of our proposed estimator.

The structure of the rest of the paper is as follows. Section 2 analyzes the process that ODS sampling data is fitted to additive hazard model with covariate adjustment. Then, in Section 3, we describe the large sample properties of our proposed estimator in progress to verify the finite sample performance of the proposed method by numerical approach in Section 4. In Section 5, the empirical research shows that the method we proposed has good practicality in the practical example of datasets from a pulmonary exacerbations analysis. Finally, the conclusion and prospect are summarized in Section 6.

2. Estimation Setup

Suppose that a cohort contains N independent subjects. For the i th ($i = 1, \dots, N$) subject, $\bar{T}_i$ is the failure time and C_i is the censoring time. $T_i = \min(\bar{T}_i, C_i)$ is the observed time and $\Delta_i = I(\bar{T}_i \leq C_i)$ is the indicator variable of the right censoring. Denote $Y_i(t) = I(T_i \geq t)$, $V_i(t) = \Delta_i I(T_i \leq t)$ and Z_i to be the at-risk process, the counting process, and the time-independent p -dimensional exposure variable, respectively. Denote τ to be the study end time.

The additive hazards model proposed by Lin and Ying [26] is as follows:

$$\lambda(t|Z_i) = \lambda_0(t) + \beta^T Z_i, \quad (2)$$

where $\lambda_0(t)$ is the unknown baseline hazard function and β is the parameter of p -dimension. If we have access to gather information about everyone's exposure, the following estimation function is commonly used for the inference of β :

$$U(\beta) = \frac{1}{N} \sum_{i=1}^N \int_0^\tau \{Z_i - \bar{Z}\} \{dV_i(t) - Y_i(t)\beta^T Z_i dt\}, \quad (3)$$

where $\bar{Z} = \sum_{i=1}^N Y(t)Z_i / \sum_{i=1}^N Y_i(t)$.

Under the ODS design, the scope of the failure time of the cases is divided into $\bar{K}$ disjoint strata $\{\bar{A}_l = (a_{l-1}, a_l]: l = 1, \dots, \bar{K}\}$ by positive constants $\{a_l: l = 1, \dots, \bar{K}\}$ satisfying $0 = a_0 < \dots < a_{l-1} < a_l < \dots < a_{\bar{K}} = \tau$. We first sample n_0 SRS individuals from the cohort, and let ξ_i be the indicator, by value 1 if the i th subject being into the SRS and 0 otherwise. Denote $\hat{\alpha} = P(\xi_i = 1) = n_0/N$. We sample subset A_k ($K \leq \bar{K}$) stratum from set $\bar{A}_l$ ($l = 1, \dots, \bar{K}$) and then n_k ($k = 1, \dots, K$) additional samples are drawn from the members who experience failure and not in the SRS, but in stratum A_k . Denote η_{ik} to be the indicator whether the i th individual from A_k is sampled into the additional samples. Denote $\hat{\gamma}_k = P(\eta_{ik} = 1 | \Delta_i = 1, \xi_i = 0, T_i \in A_k) = n_k / (\tilde{n}_k - n_{0,k})$, where $\tilde{n}_k$ and $n_{0,k}$ is the number of the cohort failure individuals and the SRS failure individuals dropped into A_k . The samples of n_0 SRS sample and n_k ($k = 1, \dots, K$) additional sample make up the ODS sample.

Denote Λ_0 to be the set of SRS individual and Λ_k ($k = 1, \dots, K$) to be the supplemental sample from A_k . Denote $\bar{\Lambda}$ to be the set of individuals outside ODS sample. Then, we can summarize the observed datasets obtained by the design of ODS as follows:

- (i) The ODS sample { The SRS sample : $\{T_i, \Delta_i, Z_i\}$, $i \in \Lambda_0$, The supplemental sample : $\{T_i, \Delta_i, Z_i | \Delta_i = 1, T_i \in A_k\}$, $i \in \Lambda_k$.
(ii) The nonvalidation sample: $\{T_i, \Delta_i\}$, $i \in \bar{\Lambda}$.

For the ODS design, we only observe the variable Z for the selected subjects. The regression parameters β can be derived by $U_w(\beta) = 0$, where

$$U_w(\beta) = \frac{1}{N} \sum_{i=1}^N \int_0^\tau \omega_i \{Z_i - \bar{Z}_w\} \{dV_i(t) - Y_i(t)\beta^T Z_i dt\}, \quad (4)$$

and $\bar{Z}_w = S^{(1)}(t)/S^{(0)}(t)$, $S^{(d)}(t) = (1/N) \sum_{i=1}^N \omega_i Y_i(t) Z_i^{\otimes d}$, ($d = 0, 1, 2$), with $a^{\otimes 0} = 1$, $a^{\otimes 1} = a$, and $a^{\otimes 2} = aa^T$ for a vector a ; and the weight ω_i is defined by

$$\omega_i = \frac{\xi_i(1 - \Delta_i)}{\hat{\alpha}} + \frac{\xi_i \Delta_i(1 - \zeta_i)}{\hat{\alpha}} + \xi_i \Delta_i \zeta_i + (1 - \xi_i) \Delta_i \sum_{k=1}^K \frac{\zeta_{ik} \eta_{ik}}{\hat{\gamma}_k}, \quad i = 1, 2, \dots, N, \quad (5)$$

where $\zeta_i = \sum_{k=1}^K \zeta_{ik}$ and $\zeta_{ik} = I(\tilde{T}_i \in A_k)$.

Note that the weight of nonvalidation samples are 0, whereas the subcohort censored individuals are $\hat{\alpha}^{-1}$. The weight of the subcohort cases are 1 if their failure time belongs to A_k ($k = 1, \dots, K$) and are $\hat{\alpha}^{-1}$, otherwise. The selected cases, not in the subcohort, are weighted by $\hat{\gamma}_k^{-1}$, when their time of failure belongs to A_k ($k = 1, \dots, K$). The estimator $\hat{\beta}$ defined by (4) takes the explicit form as follows:

$$\hat{\beta} = G_N^{-1} F_N, \quad (6)$$

where

$$G_N = \int_0^\tau \left\{ S^{(2)}(t) - \frac{S^{(1)}(t)^{\otimes 2}}{S^{(0)}(t)} \right\} dt, \quad (7)$$

$$F_N = \frac{1}{N} \sum_{i=1}^N \int_0^\tau \omega_i \{Z_i - \bar{Z}_w\} dV_i(t).$$

In practice, some covariates may be contaminated by some distorting factors. In this paper, we assume $Z_i = (H_i^T, X_i^T)^T$, where H_i is the observed $(p-1)$ -dimensional covariate and X_i is the unobservable 1-dimensional covariate and satisfies

$$W_i = \phi(U_i)X_i, \quad i = 1, 2, \dots, N, \quad (8)$$

where W_i is the actual observable 1-dimensional variable, U_i is the known confounder covariate, and $\phi(\cdot)$ is the unknown distorting function of observable variable U_i . At this point, for the ODS design, the available data have the form:

- (i) The ODS sample { The SRS sample : $\{T_i, \Delta_i, U_i, (H_i^T, W_i^T)^T\}$, $i \in \Lambda_0$, The supplemental sample : $\{T_i, \Delta_i, U_i, (H_i^T, W_i^T)^T | \Delta_i = 1, T_i \in A_k\}$, $i \in \Lambda_k$.

- (ii) The nonvalidation sample: $\{T_i, \Delta_i\}$, $i \in \bar{\Lambda}$.

Combining model (2) and equation (8), we assume $\tilde{T}_i$ ($i = 1, 2, \dots, N$) is generalized from the covariate-adjusted additive hazards model in this paper:

$$\begin{cases} \lambda(t | Z_i) = \lambda_0(t) + \theta^T H_i + \gamma X_i, \\ W_i = \phi(U_i)X_i, \end{cases} \quad (9)$$

where θ and γ are $(p-1)$ -dimensional and 1-dimensional regression parameters of primary interest, respectively. According to Şentürk and Müller [18], two conditions on model (9) can be listed as follows:

$$(C1) \ E[\phi(U)] = 1$$

$$(C2) \ \{X, U\} \text{ are mutually independent}$$

Note that condition (C1) ensured that the mean distorting effect vanishes. Based on conditions (C1) and (C2), we obtain $E[W] = E[X]$. Owing to the presence of distortion, the covariate X_i is unobservable, and the estimating function (4) can be no longer used for the inference of β . If we use directly W_i instead of X_i , it might lead to inaccurate statistical inference. Therefore, we should calibrate the covariate X_i based on the known covariate W_i and confounder covariate U_i . From (8) and condition (C2), it can obtain that

$$\phi(u) = \frac{E[W | U = u]}{E[X]}. \quad (10)$$

Define $\Phi(u) = E[W | U = u]$, and we adopt kernel method to estimate $\Phi(u)$:

$$\hat{\Phi}(u) = \frac{\sum_{i=1}^n K((u - U_i)/h) W_i}{\sum_{i=1}^n K((u - U_i)/h)}, \quad (11)$$

where $K(\cdot)$ is a kernel function and h is a bandwidth. It is easy to show that $\bar{W} = n^{-1} \sum_{i=1}^n W_i$ converges almost to $E[W]$. By $E[W] = E[X]$, equations (10) and (11), the distorting function $\phi(u)$ can be estimated by

$$\hat{\phi}(u) = \frac{\hat{\Phi}(u)}{\bar{W}}, \quad (12)$$

and the covariate X_i can be calibrated by

$$\hat{X}_i = \frac{W_i}{\hat{\phi}(U_i)}, \quad i = 1, 2, \dots, N. \quad (13)$$

Denote $\hat{Z}_i = (H_i^T, \hat{X}_i^T)^T$, and the proposed estimator $\hat{\beta}_p$ for model (9) can be defined as the solution of the function:

$$\hat{U}_w(\beta) = \frac{1}{N} \sum_{i=1}^N \int_0^\tau \omega_i \{ \hat{Z}_i - \hat{\bar{Z}}_w \} \{dV_i(t) - Y_i(t)\beta^T \hat{Z}_i dt\}, \quad (14)$$

where $\hat{\bar{Z}}_w = \hat{S}^{(1)}(t)/\hat{S}^{(0)}(t)$ and $\hat{S}^{(d)}(t) = (1/N) \sum_{i=1}^N \omega_i Y_i(t) \hat{Z}_i^{\otimes d}$, ($d = 0, 1, 2$). The explicit form of the proposed estimator $\hat{\beta}_p$ can be described as follows by some simple calculation:

$$\hat{\beta}_p = \hat{G}_N^{-1} \hat{F}_N, \quad (15)$$

where

$$\hat{G}_N = \int_0^\tau \left\{ \hat{S}^{(2)}(t) - \frac{\hat{S}^{(1)}(t)^{\otimes 2}}{\hat{S}^{(0)}(t)} \right\} dt, \quad (16)$$

$$\hat{F}_N = \frac{1}{N} \sum_{i=1}^N \int_0^\tau w_i \left\{ \hat{Z}_i - \hat{\bar{Z}}_\omega \right\} dV_i(t).$$

3. Main Results

In this section, we would like to establish the asymptotic properties of $\hat{\beta}_p$ in (15).

Firstly, we give the following definition.

Definition 1. Define

$$s^{(0)}(t) = E[Y_1(t)],$$

$$s_{HX} = E[Y_1(t)HX],$$

$$s_X^{(d)}(t) = E[Y_1(t)X^{\otimes d}],$$

$$s_H^{(d)}(t) = E[Y_1(t)H^{\otimes d}], \quad d = 1, 2,$$

$$M_i(t) = V_i(t) - \int_0^t Y_i(s)\lambda_0(s)ds - \int_0^t Y_i(s)\beta^T Z_i ds, \quad (17)$$

$$S_i(\beta_0) = \int_0^\tau \left\{ Z_i - \frac{E[Y_1(t)Z]}{E[Y_1(t)]} \right\} dM_i(t), \quad i = 1, 2, \dots, N,$$

$$D_1(\beta_0) = E[(1 - \Delta_1)S_1(\beta_0)^{\otimes 2}],$$

$$D_2(\beta_0) = E[\Delta_1(1 - \zeta_1)S_1(\beta_0)^{\otimes 2}],$$

$$D_3(\beta_0) = E[\Delta_1\zeta_{1k}S_1(\beta_0)^{\otimes 2}],$$

$$\sum (\beta_0) = E \left[\int_0^\tau \left\{ Z_1 - \frac{E[Y_1(t)Z]}{E[Y_1(t)]} \right\}^{\otimes 2} dV_1(t) \right], \quad (18)$$

$$\begin{aligned} \sum_1 (\beta_0) &= \frac{1-\alpha}{\alpha} D_1(\beta_0) + \frac{1-\alpha}{\alpha} D_2(\beta_0) \\ &+ \sum_{k=1}^K (1-\alpha) \frac{1-\gamma_k}{\gamma_k} D_3(\beta_0), \end{aligned} \quad (19)$$

$$G_1 = \int_0^\tau \left\{ s_H^{(2)}(t) - \frac{s_H^{(1)}(t)^{\otimes 2}}{s^{(0)}(t)} \right\} dt, \quad (20)$$

$$G_2 = \int_0^\tau \left\{ s_{HX}(t) - \frac{s_H^{(1)}(t)s_X^{(1)}(t)}{s^{(0)}(t)} \right\} dt, \quad (21)$$

$$G_3 = \int_0^\tau \left\{ s_X^{(2)}(t) - \frac{s_X^{(1)}(t)^2}{s^{(0)}(t)} \right\} dt, \quad (22)$$

where $M_i(t)$ is a locally square integrable martingale (Lin and Ying [26]) and $\beta_0 = (\theta_0^T, \gamma_0^T)$ is the true parameters value.

Secondly, the following additional regular conditions are concluded to illustrate the process:

$$(C3) \int_0^\tau \lambda_0(t) dt < \infty.$$

$$(C4) P(Y_1(t) = 1, \text{ for any } t \in [0, \tau]) > 0.$$

$$(C5) E[\sup_{0 \leq t \leq \tau} |Y_1(t)Z_1^{\otimes 2}(\beta^T Z_1^2)|] < \infty.$$

$$(C6) \text{The matrix } \sum (\beta_0) \text{ appeared in (18) is finite positive definite.}$$

$$(C7) \text{The matrix } G = \begin{bmatrix} G_1 & G_2 \\ G_2^T & G_3 \end{bmatrix} \text{ is nonsingular.}$$

$$(C8) \text{As } N \longrightarrow \infty, \quad \hat{\alpha} \longrightarrow \alpha \in (0, 1) \quad \text{and} \quad \hat{\gamma}_k \longrightarrow \gamma_k \in (0, 1) \text{ for } k = 1, \dots, K.$$

$$(C9) \phi(u), f(u), \text{ and } g(u) = \phi(u)f(u) \text{ are differential, and the 3-order derivatives of } f(u) \text{ and } g(u) \text{ meet the following condition, and there exist } a > 0 \text{ and a neighborhood of origin, such that, if } \delta \text{ fails to the neighborhood, we have } |f^{(3)}(u + \delta) - f^{(3)}(u)| \leq a|\delta| \text{ and } |g^{(3)}(u + \delta) - g^{(3)}(u)| \leq a|\delta|, \text{ where } f(u) \text{ is a density function of } U.$$

$$(C10) \text{The function } K(\cdot) \text{ appeared in (11) meets}$$

$$(i) \text{The support of } K(\cdot) \text{ is the interval } [-1, 1].$$

$$(ii) K(\cdot) \text{ is symmetric about zero.}$$

$$(iii) \int_{-1}^1 K(u) du = 1, \quad \int_{-1}^1 u^i K(u) du = 0, \quad i = 1, 2, 3.$$

$$(C11) \text{As } N \longrightarrow \infty, \text{ the bandwidth } h \text{ falls from } O(N^{-1/4} \log N) \text{ to } O(N^{-1/8}).$$

$$(C12) E[X] \text{ is bounded away from 0 and } E[X]^2 < \infty.$$

The above conditions are mild and suitable in many circumstances. Conditions (C3)–(C8) are regular conditions of the regression parameters which are similar to Yu et al. [27]. However, the likes of conditions (C9)–(C12) can be traced to Cui et al. [20].

Theorem 1. *If conditions (C1)–(C12) hold, as $N \longrightarrow \infty$, we have $\hat{\beta}_p \longrightarrow \beta_0$.*

Theorem 2. *If conditions (C1)–(C12) hold, as $N \longrightarrow \infty$, we have*

$$\sqrt{N}(\hat{\beta}_p - \beta_0) \xrightarrow{d} \mathcal{N}\left(0, G^{-1}\left(\sum(\beta_0) + \sum_1(\beta_0) + \gamma_0^2 \frac{\text{Var}(W) - (3/4)\text{Var}(X)}{[\text{E}(X)]^2} LL^T\right)G^{-1}\right), \quad (23)$$

where $L = [G_2, G_3]^T$. In other words,

$$\sqrt{N}(\hat{\theta}_p - \theta_0) \xrightarrow{d} \mathcal{N}\left(0, G^{-1}\left(\sum(\beta_0) + \sum_1(\beta_0)\right)G^{-1}\right), \sqrt{N}(\hat{\gamma}_p - \gamma_0) \xrightarrow{d} \mathcal{N}\left(0, \sigma + \gamma_0^2 \frac{\text{Var}(W) - 3/4\text{Var}(X)}{[\text{E}(X)]^2}\right), \quad (24)$$

where $\hat{\beta}_p = (\hat{\theta}_p^T, \hat{\gamma}_p^T)^T$ and σ is the $(p+1, p+1)$ -element of matrix $G^{-1}(\sum(\beta_0) + \sum_1(\beta_0))G^{-1}$.

To prove Theorem 1, the following definition and deformation lemmas are needed.

Definition 2. Define

$$\begin{aligned} S_H^{(d)}(t) &= \frac{1}{N} \sum_{i=1}^N \omega_i Y_i(t) H_i^{\otimes d}, \\ S_X^{(d)}(t) &= \frac{1}{N} \sum_{i=1}^N \omega_i Y_i(t) X_i^{\otimes d}, \\ S_{\hat{X}}^{(d)}(t) &= \frac{1}{N} \sum_{i=1}^N \omega_i Y_i(t) \hat{X}_i^{\otimes d}, \quad d = 1, 2, \\ S_{HX}(t) &= \frac{1}{N} \sum_{i=1}^N \omega_i Y_i(t) H_i X_i, \\ S_{H\hat{X}}(t) &= \frac{1}{N} \sum_{i=1}^N \omega_i Y_i(t) H_i \hat{X}_i. \end{aligned} \quad (25)$$

In order to prove conveniently the results, we define the partitioned matrices:

$$\begin{aligned} \hat{F}_N &= \begin{bmatrix} \hat{F}_{N_1} \\ \hat{F}_{N_2} \end{bmatrix}, \\ F_N &= \begin{bmatrix} F_{N_1} \\ F_{N_2} \end{bmatrix}, \\ \hat{G}_N &= \begin{bmatrix} \hat{G}_{N_1} & \hat{G}_{N_2} \\ \hat{G}_{N_2}^T & \hat{G}_{N_3} \end{bmatrix}, \\ G_N &= \begin{bmatrix} G_{N_1} & G_{N_2} \\ G_{N_2}^T & G_{N_3} \end{bmatrix}, \end{aligned} \quad (26)$$

where

$$\begin{aligned} \hat{G}_{N_1} &= G_{N_1} = \int_0^\tau \left[S_H^{(2)}(t) - \frac{S_H^{(1)}(t)^{\otimes 2}}{S^{(0)}(t)} \right] dt, \\ \hat{G}_{N_2} &= \int_0^\tau \left[S_{H\hat{X}}(t) - \frac{S_H^{(1)}(t) S_{\hat{X}}^{(1)}(t)}{S^{(0)}(t)} \right] dt, \\ G_{N_2} &= \int_0^\tau \left[S_{HX}(t) - \frac{S_H^{(1)}(t) S_X^{(1)}(t)}{S^{(0)}(t)} \right] dt, \\ \hat{G}_{N_3} &= \int_0^\tau \left[S_{\hat{X}}^{(2)}(t) - \frac{S_{\hat{X}}^{(1)}(t)^2}{S^{(0)}(t)} \right] dt, \\ G_{N_3} &= \int_0^\tau \left[S_X^{(2)}(t) - \frac{S_X^{(1)}(t)^2}{S^{(0)}(t)} \right] dt, \\ \hat{F}_{N_1} &= F_{N_1} = \frac{1}{N} \sum_{i=1}^N \int_0^\tau \omega_i \left[H_i - \frac{S_H^{(1)}(t)}{S^{(0)}(t)} \right] dV_i(t), \\ \hat{F}_{N_2} &= \frac{1}{N} \sum_{i=1}^N \int_0^\tau \omega_i \left[\hat{X}_i - \frac{S_{\hat{X}}^{(1)}(t)}{S^{(0)}(t)} \right] dV_i(t), \\ F_{N_2} &= \frac{1}{N} \sum_{i=1}^N \int_0^\tau \omega_i \left[X_i - \frac{S_X^{(1)}(t)}{S^{(0)}(t)} \right] dV_i(t). \end{aligned} \quad (27)$$

Lemma 1. If conditions (C3)–(C8) hold, then

$$\sup_{t \in [0, \tau]} \|S^{(d)}(t) - s^{(d)}(t)\| \xrightarrow{P} 0, \quad d = 0, 1, 2. \quad (28)$$

Proof. Applying Glivenko–Cantelli theorem, we obtain that

$$\sup_{t \in [0, \tau]} \|S^{(0)}(t) - s^{(0)}(t)\| \xrightarrow{P} 0, \quad (29)$$

where $s^{(0)}(t) = \text{E}[Y_1(t)]$. By corollary (III).2 from Andersen and Gill [28], the uniform convergence of $S^{(1)}(t)$ and $S^{(2)}(t)$ can be similarly shown. $\square$

Lemma 2 (see from Cui et al. [20]). *If conditions (C1), (C2), and (C9)–(C12) hold, then*

$$\frac{1}{N} \sum_{i=1}^N (\hat{X}_i - X_i) \psi(H_i, X_i, t) - Q_N E[X \psi(H, X, t)] = o_p(N^{-1/2}), \quad (30)$$

where $\psi(h, x, t)$ is a function of h , x , and t satisfying $E[\psi(H, X, t)]^2 < \infty$, and

$$Q_N = \frac{1}{NE[X]} \left[\sum_{i=1}^N \left(W_i - \frac{X_i}{2} - \frac{E[X]}{2} \right) \right]. \quad (31)$$

Lemma 3. *If conditions (C1)–(C12) hold, then*

$$\begin{aligned} \hat{F}_N - F_N &= o_p(1), \\ \hat{G}_N - G_N &= o_p(1). \end{aligned} \quad (32)$$

Proof. By Lemma 2 with $\psi(H_i, X_i, t) = \omega_i Y_i(t)$ and $E[Q_N] = 0$, we obtain

$$\begin{aligned} S_X^{(1)}(t) - S_X^{(1)}(t) &= \frac{1}{N} \sum_{i=1}^N [\omega_i Y_i(t) (\hat{X}_i - X_i)] \\ &= s_X^{(1)}(t) Q_N + o_p(N^{-1/2}) = o_p(1), \end{aligned} \quad (33)$$

where $s_X^{(1)}(t) = E[Y_1(t)X]$. By Lemma 2 with $\psi(H_i, X_i, t) = \omega_i Y_i(t)H_i$, we have

$$\begin{aligned} S_{H\hat{X}}(t) - S_{HX}(t) &= \frac{1}{N} \sum_{i=1}^N \omega_i Y_i(t) H_i (\hat{X}_i - X_i) \\ &= s_{HX}(t) Q_N + o_p(N^{-1/2}) = o_p(1), \end{aligned} \quad (34)$$

where $s_{HX}(t) = E[Y_1(t)HX]$. Applying Lemma 2 with $\psi(H_i, X_i, t) = \omega_i Y_i(t)X_i$, we obtain

$$\begin{aligned} S_X^{(2)}(t) - S_X^{(2)}(t) &= \frac{1}{N} \sum_{i=1}^N \omega_i Y_i(t) (\hat{X}_i^2 - X_i^2) \\ &= \frac{2}{N} \sum_{i=1}^N \omega_i Y_i(t) X_i (\hat{X}_i - X_i) \\ &\quad + \frac{1}{N} \sum_{i=1}^N \omega_i Y_i(t) (\hat{X}_i - X_i)^2 \\ &= 2s_X^{(2)}(t) Q_N + \frac{1}{N} \sum_{i=1}^N \omega_i Y_i(t) (\hat{X}_i - X_i)^2 \\ &\quad + o_p(N^{-1/2}), \end{aligned} \quad (35)$$

where $s_X^{(2)}(t) = E[Y_1(t)X^{\otimes 2}]$. Similar to the proof of $D_2 = o_p(N^{1/2})$ in Proposition 1 from Cui et al. [20], we have

$$\sup_{0 \leq t \leq \tau} \sum_{i=1}^N [\omega_i Y_i(t) (\hat{X}_i - X_i)^2] = o_p(N^{1/2}). \quad (36)$$

Therefore,

$$S_X^{(2)}(t) - S_X^{(2)}(t) = 2s_X^{(2)}(t) Q_N + o_p(N^{-1/2}) = o_p(1). \quad (37)$$

By equation (33) and the uniform convergence of $S_X^{(1)}(t)$ to $s_X^{(1)}(t)$, we obtain

$$\begin{aligned} S_X^{(1)}(t)^2 - S_X^{(1)}(t)^2 &= (S_X^{(1)}(t) - S_X^{(1)}(t)) (S_X^{(1)}(t) + S_X^{(1)}(t)) \\ &= (s_X^{(1)}(t) Q_N + o_p(N^{-1/2})) \\ &\quad \cdot (2S_X^{(1)}(t) + o_p(1)) \\ &= 2s_X^{(1)}(t)^2 Q_N + o_p(N^{-1/2}). \end{aligned} \quad (38)$$

From the partitioned matrices in the above notation, we obtain that

$$\begin{aligned} \hat{F}_N - F_N &= \begin{bmatrix} 0 \\ \hat{F}_{N_1} - F_{N_2} \end{bmatrix}, \\ \hat{G}_N - G_N &= \begin{bmatrix} 0 & \hat{G}_{N_2} - G_{N_2} \\ \hat{G}_{N_2}^T - G_{N_2}^T & \hat{G}_{N_3} - G_{N_3} \end{bmatrix}. \end{aligned} \quad (39)$$

Then, by equation (33) and Lemma 2 with $\psi(H_i, X_i, t) = V_i(\tau)$, we obtain that

$$\begin{aligned} \hat{F}_{N_2} - F_{N_2} &= \frac{1}{N} \sum_{i=1}^N [\omega_i (\hat{X}_i - X_i) V_i(\tau)] \\ &\quad - \frac{1}{N} \sum_{i=1}^N \omega_i \int_0^\tau \left\{ \frac{S_X^{(1)}(t) - S_X^{(1)}(t)}{S^{(0)}(t)} \right\} dV_i(t) \\ &= Q_N E[XV_1(\tau)] - \frac{Q_N}{N} \sum_{i=1}^N \omega_i \int_0^\tau \frac{s_X^{(1)}(t)}{s^{(0)}(t)} dV_i(t) \\ &\quad + o_p(N^{-1/2}) = Q_N E[XV_1(\tau)] \\ &\quad - Q_N E \left[\int_0^\tau \frac{s_X^{(1)}(t)}{s^{(0)}(t)} dV_1(t) \right] + o_p(N^{-1/2}) \\ &= Q_N E \left[\int_0^\tau \left\{ X - \frac{s_X^{(1)}(t)}{s^{(0)}(t)} \right\} dV_1(t) \right] + o_p(N^{-1/2}) \\ &= o_p(1), \end{aligned} \quad (40)$$

where $s^{(0)}(t) = E[Y_1(t)]$. By equations (39) and (40), we have

$$\hat{F}_N = F_N + o_p(1). \quad (41)$$

By equations (33) and (34), we have

$$\begin{aligned}\widehat{G}_{N_2} - G_{N_2} &= \int_0^\tau \left[\left(S_{HX}(t) - S_{HX}(t) \right) - \frac{S_H^{(1)}(t) \left(S_X^{(1)}(t) - S_X^{(1)}(t) \right)}{S^{(0)}(t)} \right] dt \\ &= \int_0^\tau \left[s_{HX}(t) Q_N - \frac{s_H^{(1)}(t) s_X^{(1)}(t)}{s^{(0)}(t)} Q_N \right] dt = Q_N G_2 + o_p(N^{-1/2}) = o_p(1),\end{aligned}\quad (42)$$

where G_2 appeared in (21). By equations (37) and (38), we have

$$\begin{aligned}\widehat{G}_{N_3} - G_{N_3} &= \int_0^\tau \left[S_X^{(2)}(t) - S_X^{(2)}(t) - \frac{S_X^{(1)}(t)^2 - S_X^{(1)}(t)^2}{S^{(0)}(t)} \right] dt \\ &= \int_0^\tau \left[2s_X^{(2)}(t) Q_N - \frac{2s_X^{(1)}(t)^2 Q_N}{s^{(0)}(t)} \right] dt + o_p(N^{-1/2}) = 2Q_N G_3 + o_p(N^{-1/2}) = o_p(1),\end{aligned}\quad (43)$$

where G_3 appeared in (22). Then, by equations (39), (42), and (43), we have

$$\widehat{G}_N = G_N + o_p(1). \quad (44)$$

Proof of Theorem 1. Applying Lemma 3 and the definition of the notation of G_N and G , we obtain

$$\widehat{G}_N \xrightarrow{P} G. \quad (45)$$

By equation (45), condition (C7) and Slutsky Lemma, we show that $\widehat{G}_N$ is nonsingular in probability. Thus, it obtains that

$$\begin{aligned}\widehat{\beta} &= G_N^{-1} F_N, \\ \widehat{\beta}_p &= \widehat{G}_N^{-1} \widehat{F}_N.\end{aligned}\quad (46)$$

Similar to Yu et al. [27], we get $\widehat{\beta} \xrightarrow{P} \beta_0$; then, by equations (45) and (46), Lemma 3 and Slutsky Lemma, it holds that

$$\widehat{\beta}_p \xrightarrow{P} \beta_0. \quad (47)$$

Therefore, Theorem 1 has been proved. $\square$

To prove Theorem 2, the following deformation lemma is needed.

Lemma 4. *If conditions (C1)–(C12) hold, then*

$$\widehat{U}_\omega(\beta_0) = U_\omega(\beta_0) - \gamma_0 Q_N L + o_p(N^{-1/2}), \quad (48)$$

where $L = [G_2, G_3]^T$, G_2 and G_3 appeared in (21) and (22), respectively.

Proof. By equations (4) and (14), we obtain

$$\begin{aligned}\widehat{U}_\omega(\beta_0) &= \widehat{F}_N - \widehat{G}_N \beta_0 = \begin{bmatrix} \widehat{F}_{N_1} - \widehat{G}_{N_1} \theta_0 - \widehat{G}_{N_2} \gamma_0 \\ \widehat{F}_{N_2} - \widehat{G}_{N_2}^T \theta_0 - \widehat{G}_{N_3} \gamma_0 \end{bmatrix}, \\ U_\omega(\beta_0) &= F_N - G_N \beta_0 = \begin{bmatrix} F_{N_1} - G_{N_1} \theta_0 - G_{N_2} \gamma_0 \\ F_{N_2} - G_{N_2}^T \theta_0 - G_{N_3} \gamma_0 \end{bmatrix}.\end{aligned}\quad (49)$$

By equation (41), we have

$$\widehat{F}_{N_2} - F_{N_2} = Q_N E \left[\int_0^\tau \left\{ X - \frac{s_X^{(1)}(t)}{s^{(0)}(t)} \right\} dV_1(t) \right] + o_p(N^{-1/2}). \quad (50)$$

By equations (42) and (43), we have

$$\begin{aligned}\widehat{G}_{N_2} - G_{N_2} &= Q_N G_2 + o_p(N^{-1/2}), \\ \widehat{G}_{N_3} - G_{N_3} &= 2Q_N G_3 + o_p(N^{-1/2}).\end{aligned}\quad (51)$$

Similar to Yu et al. [27], we have

$$U_\omega(\beta_0) = \frac{1}{N} \sum_{i=1}^N \int_0^\tau \omega_i \left\{ Z_i - \frac{S^{(1)}(t)}{S^{(0)}(t)} \right\} dM_i(t), \quad (52)$$

where $M_i(t)$ appeared in (17), which is a square integrable martingale. Then, performing a simple calculation, we obtain that

$$\widehat{U}_\omega(\beta_0) - U_\omega(\beta_0) = \begin{bmatrix} (G_{N_2} - \widehat{G}_{N_2}) \gamma_0 \\ \widehat{F}_{N_2} - F_{N_2} + (G_{N_2} - \widehat{G}_{N_2})^T \theta_0 + (G_{N_3} - \widehat{G}_{N_3}) \gamma_0 \end{bmatrix}. \quad (53)$$

By $U_\omega(\beta_0) = F_N - G_N\beta_0$ and $E[U_\omega(\beta_0)] = 0$, we have

$$E[F_N - G_N\beta_0] = 0. \quad (54)$$

Then, we have $E[F_{N2}] = (E[G_{N2}])^T \theta_0 + E[G_{N3}]\gamma_0$, that is,

$$E\left[\int_0^\tau \left\{X - \frac{s_X^{(1)}(t)}{s^{(0)}(t)}\right\} dV_1(t)\right] = G_2^T \theta_0 + G_3 \gamma_0. \quad (55)$$

By equations (50) and (55), we have

$$\hat{F}_{N_2} - F_{N_2} = Q_N(G_2^T \theta_0 + G_3 \gamma_0) + o_p(N^{-1/2}). \quad (56)$$

Therefore, by equations (51), (53), and (56), we have

$$\begin{aligned} \hat{U}_\omega(\beta_0) &= U_\omega(\beta_0) - \gamma_0 Q_N[G_2, G_3]^T + o_p(N^{-1/2}) \\ &= U_\omega(\beta_0) - \gamma_0 Q_N L + o_p(N^{-1/2}). \end{aligned} \quad (57)$$

Theorem 2 can be proven by using the properties of Lemma 4. $\square$

Proof of Theorem 2. By Lemma 4, we have

$$\sqrt{N}\hat{U}_\omega(\beta_0) = \sqrt{N}U_\omega(\beta_0) - \sqrt{N}\gamma_0 Q_N L + o_p(1). \quad (58)$$

Through the calculation of double expectation, we have

$$\begin{aligned} \text{Cov}(\sqrt{N}U_\omega(\beta_0), \sqrt{N}Q_N) &= \frac{1}{E[X]} \\ E\left\{U_\omega(\beta_0) \left[\sum_{i=1}^N \left(W_i - \frac{X_i}{2} - \frac{E[X]}{2}\right) \middle| H_i, W_i, U_i\right]\right\} &= 0. \end{aligned} \quad (59)$$

By Yu et al. [27], we have

$$\text{Var}[\sqrt{N}U_\omega(\beta_0)] = \sum (\beta_0) + \sum_1 (\beta_0), \quad (60)$$

where $\sum (\beta_0)$ and $\sum_1 (\beta_0)$ appeared in (20) and (21), respectively. By performing a simple calculation, we obtain

$$\begin{aligned} \text{Var}(\sqrt{N}Q_N) &= \frac{1}{(E[X])^2} \text{Var}\left(W - \frac{X}{2}\right) \\ &= \frac{1}{(E[X])^2} \left[\text{Var}(W) + \frac{1}{4} \text{Var}(X) - \text{Cov}\left(W, \frac{X}{2}\right) \right]. \end{aligned} \quad (61)$$

By conditions (C1) and (C2), we have

$$\text{Cov}\left(W, \frac{X}{2}\right) = E(\phi(U))E(X^2) - E(\phi(U))E(X)^2 = \text{Var}(X). \quad (62)$$

Combining equations (61) and (62), we obtain that

$$\text{Var}(\sqrt{N}Q_N) = \frac{\text{Var}(W) - (3/4)\text{Var}(X)}{[E(X)]^2}. \quad (63)$$

Thus, by (58)–(60) and (63), we have

$$\begin{aligned} \text{Var}(\sqrt{N}\hat{U}_\omega(\beta_0)) &\xrightarrow{P} \sum (\beta_0) + \sum_1 (\beta_0) \\ &\quad + \gamma_0^2 \frac{\text{Var}(W) - (3/4)\text{Var}(X)}{(E(X))^2} LL^T. \end{aligned} \quad (64)$$

Then, we obtain that

$$\begin{aligned} \sqrt{N}(\hat{\beta}_p - \beta_0) &= \left[\frac{\partial \hat{U}_\omega(\beta)}{\partial \beta} \middle|_{\beta=\bar{\beta}_n} \right]^{-1} [\sqrt{N}\hat{U}_\omega(\beta_0)] = -G^{-1} [\sqrt{N}\hat{U}_\omega(\beta_0)] \\ &\quad + o_p(1) \xrightarrow{d} \mathcal{N}\left(0, G^{-1} \left(\sum (\beta_0) + \sum_1 (\beta_0) + \gamma_0^2 \frac{\text{Var}(W) - (3/4)\text{Var}(X)}{[E(X)]^2} LL^T \right) G^{-1} \right), \end{aligned} \quad (65)$$

where $\bar{\beta}_n$ is a point of the line between $\hat{\beta}_p$ and β_0 , and

$$\begin{aligned} &G^{-1} \left(\sum (\beta_0) + \sum_1 (\beta_0) + \gamma_0^2 \frac{\text{Var}(W) - (3/4)\text{Var}(X)}{[E(X)]^2} LL^T \right) G^{-1} \\ &= G^{-1} \left(\sum (\beta_0) + \sum_1 (\beta_0) \right) G^{-1} + \gamma_0^2 \frac{\text{Var}(W) - (3/4)\text{Var}(X)}{[E(X)]^2} G^{-1} (G_2, G_3)^T (G_2^T, G_3) G^{-1} \\ &= G^{-1} \left(\sum (\beta_0) + \sum_1 (\beta_0) \right) G^{-1} + \gamma_0^2 \frac{\text{Var}(W) - (3/4)\text{Var}(X)}{[E(X)]^2} R, \end{aligned} \quad (66)$$

where R is a $(p+1)$ -dimensional matrix with all elements being zero except for the $(p+1) \times (p+1)$ -element being 1. Therefore, we conclude that

$$\sqrt{N}(\hat{\theta}_p - \theta_0) \xrightarrow{d} \mathcal{N}\left(0, G^{-1}\left(\sum (\beta_0) + \sum_1 (\beta_0)\right)G^{-1}\right), \sqrt{N}(\hat{\gamma}_p - \gamma_0) \xrightarrow{d} \mathcal{N}\left(0, \sigma + \gamma_0^2 \frac{\text{Var}(W) - (3/4)\text{Var}(X)}{[E(X)]^2}\right), \quad (67)$$

where σ is the $(p+1, p+1)$ -element of matrix $G^{-1}(\sum (\beta_0) + \sum_1 (\beta_0))G^{-1}$.

Here, Theorem 2 has been proven regarding the asymptotic convergence properties of normal distribution. It was from a different viewpoint compared to previous research. $\square$

4. Numerical Approach

Strictly speaking, we carry out some simulations in the section. The underlying additive hazards model considered is as follows:

$$\lambda(t|H, X) = \lambda_0(t) + \theta H + \gamma X, \quad (68)$$

where the baseline function $\lambda_0(t)$ is set to be $3t^2 + 1$ and $e^{-t} + t$, respectively. The true parameters $\theta = 0.3$ and $\gamma = -0.2$. The covariate $H \sim \mathcal{N}(0, 1)$ and $X \sim E(1)$. The censoring time $C \sim \text{uniform}(0, c)$, where constant c is chosen to approximately produce three censoring rate $\rho = 30\%$, 50% , and 80% . The confounding variable $U \sim \text{uniform}(4 - \sqrt{7}, 4 + \sqrt{7})$ and the distortion function $\phi(U) = (1 + U)^3/a$, where the constant a is chosen such that the distorting function satisfies the identifiability constraint $E[\phi(U)] = 1$. The observed covariate $W = \phi(U)X$. We choose a high-order kernel function $K(t) = (15/32)(3 - 7t^2)(1 - t^2)I(|t| \leq 1)$ and use leave-one-out crossvalidation to select the bandwidth.

Under the design of ODS, we sample $n_0 = 400$ subcohort individuals without replacement from $N = 4000$. Then we partition the observed failure time into three strata by quantiles of observed failure time. In order to study the influence of different cutpoints, we consider 0.2 and 0.8 quantile and 0.3 and 0.7 quantile, respectively. We sample the additional individuals of size $n_1 = 25$ and $n_3 = 25$ from the first stratum and the third stratum. In addition, we compare our proposed covariate-adjusted estimator (Proposed) with two estimators, for example, oracle estimator (Oracle) which is calculated based on the true covariate X and naive estimator (Naive) which is computed based on the contaminated covariate W . Note that the oracle estimator is computed from the observations of X , which is not available in the real data. Meanwhile, the naive method is sure to exploit to regard directly the contaminated covariate W as the true covariate X . Under each configuration, the results presented in Tables 1 and 2 are obtained from 1000 independently generated datasets, including the biases of the estimates (Bias), the sample standard deviation (SD), the

estimated standard error (SE), and the 95% normal confidence interval (CP).

By comparison and analysis, the oracle estimator is considering to be the best of all three estimators. To be specific, the proposed estimators for both θ and γ are all unbiased, and the statistical performance can compete with that of the oracle estimator. Foremost, the 95% normal confidence intervals are reasonable. When it comes to the naive estimator, the result for γ is biased. However, through covariate-adjusted process, the main results in Section 3. In addition, it is a fact that the efficiency gains are higher when cutpoint is (0.2, 0.8) than when cutpoint is (0.3, 0.7).

Additionally, we conduct simulation studies to evaluate the behavior of the proposed method when the censoring time C depends on the covariate. The setup is the same as in Tables 1 and 2, except that the censoring time C is taken as $U[0, 1]I(H < 0) + U[0, 3]I(H \geq 0)$ and $E(1)I(H < 0) + E(1/3)I(H \geq 0)$, respectively. The results are reported in Table 3 when $\lambda_0(t) = 3t^2 + 1$ and Table 4 when $\lambda_0(t) = e^{-t} + t$, which show that the proposed method performs satisfactorily in the cases.

5. Empirical Analysis

Studies have been completed to conclude the real-world analysis. Our study data contains 641 patients. The accumulation of extracellular DNA in the lung during bacterial infection can bring out progressive deterioration of lung function and aggravation of respiratory symptoms in patients with cystic fibrosis. Therefore, the dependent variable that we are interested in is time to relapse, and the censoring rate of the dependent variable is approximately 62.4%. Under the ODS design, we sample 200 individuals as subcohort sample. We partition the dependent variable that are not in the subcohort into three strata. We choose two kinds of cutpoints similar to the simulations. The supplemental samples of size $n_1 = 10$ and $n_3 = 10$ are selected from the first stratum and the third stratum.

Two variables relevant to potential confounders have been found, such as vital capacity and patient's type of treatment (Type), divided into placebo and rhDNase. In this study, we measured forced expiratory volume twice and abbreviated FEV_1 and FEV_2 separately. Then, we regarded $FEV((FEV_1 + FEV_2)/2)$ as a disorder index of vital capacity. It has become apparent that the confounder factor U follows a uniform distribution over $[0, 1]$ on the basis of average. A comprehensive study of the additive hazards model has been

TABLE 1: Simulation results when $\lambda_0(t) = 3t^2 + 1$.

ρ	Cutpoints	Method	$\hat{\theta}$				$\hat{\gamma}$			
			Bias	SD	SE	CP	Bias	SD	SE	CP
30%	(0.3, 0.7)	Oracle	-0.0034	0.1198	0.1149	0.931	0.0052	0.0991	0.0948	0.941
		Naive	-0.0032	0.1202	0.1151	0.935	0.1154	0.0697	0.0631	0.523
		Proposed	-0.0034	0.1200	0.1149	0.932	0.0014	0.1024	0.0980	0.943
	(0.2, 0.8)	Oracle	-0.0012	0.0996	0.0962	0.936	0.0014	0.0801	0.0800	0.942
		Naive	-0.0018	0.0994	0.0962	0.935	0.1119	0.0544	0.0531	0.436
		Proposed	-0.0012	0.0996	0.0962	0.937	-0.0008	0.0828	0.0830	0.945
50%	(0.3, 0.7)	Oracle	-0.0018	0.0959	0.0978	0.944	0.0030	0.0797	0.0792	0.950
		Naive	-0.0018	0.0958	0.0977	0.947	0.1138	0.0544	0.0537	0.428
		Proposed	-0.0018	0.0961	0.0978	0.942	0.0015	0.0828	0.0822	0.955
	(0.2, 0.8)	Oracle	-0.0074	0.0871	0.0880	0.953	0.0026	0.0723	0.0715	0.951
		Naive	-0.0073	0.0869	0.0880	0.956	0.1109	0.0510	0.0485	0.351
		Proposed	-0.0075	0.0869	0.0880	0.955	-0.0002	0.0753	0.0745	0.951
80%	(0.3, 0.7)	Oracle	-0.0015	0.0940	0.0951	0.952	0.0168	0.0777	0.0752	0.957
		Naive	-0.0020	0.0934	0.0951	0.950	0.1184	0.0533	0.0501	0.308
		Proposed	-0.0016	0.0939	0.0951	0.953	0.0155	0.0817	0.0783	0.959
	(0.2, 0.8)	Oracle	-0.0062	0.0941	0.0927	0.941	0.0163	0.0717	0.0725	0.949
		Naive	-0.0062	0.0934	0.0926	0.941	0.1157	0.0507	0.0486	0.304
		Proposed	-0.0063	0.0941	0.0927	0.942	0.0146	0.0750	0.0755	0.952

“Oracle” denotes the oracle estimator calculated based on the true covariate. “Naive” denotes the naive estimator calculated based on the unadjusted covariate. “Proposed” denotes the proposed estimator calculated based on the adjusted covariate.

TABLE 2: Simulation results when $\lambda_0(t) = e^{-t} + t$.

ρ	Cutpoints	Method	$\hat{\theta}$				$\hat{\gamma}$			
			Bias	SD	SE	CP	Bias	SD	SE	CP
30%	(0.3, 0.7)	Oracle	-0.0005	0.0774	0.0761	0.935	0.0071	0.0560	0.0544	0.945
		Naive	-0.0005	0.0770	0.0759	0.943	0.1131	0.0391	0.0369	0.155
		Proposed	-0.0006	0.0774	0.0760	0.932	0.0035	0.0591	0.0571	0.941
	(0.2, 0.8)	Oracle	-0.0027	0.0670	0.0645	0.931	0.0059	0.0480	0.0476	0.945
		Naive	-0.0037	0.0666	0.0645	0.928	0.1123	0.0344	0.0319	0.090
		Proposed	-0.0031	0.0670	0.0645	0.932	0.0028	0.0503	0.0499	0.947
50%	(0.3, 0.7)	Oracle	-0.0033	0.0694	0.0706	0.958	0.0110	0.0524	0.0527	0.950
		Naive	-0.0042	0.0696	0.0705	0.958	0.1142	0.0352	0.0357	0.118
		Proposed	-0.0034	0.0692	0.0706	0.961	0.0084	0.0550	0.0550	0.955
	(0.2, 0.8)	Oracle	-0.0083	0.0617	0.0640	0.962	0.0106	0.0478	0.0480	0.958
		Naive	-0.0093	0.0621	0.0640	0.959	0.1156	0.0339	0.0323	0.067
		Proposed	-0.0083	0.0618	0.0640	0.959	0.0074	0.0502	0.0501	0.958
80%	(0.3, 0.7)	Oracle	-0.0144	0.0862	0.0851	0.945	0.0169	0.0664	0.0655	0.951
		Naive	-0.0142	0.0864	0.0851	0.940	0.1177	0.0461	0.0443	0.224
		Proposed	-0.0144	0.0863	0.0851	0.946	0.0146	0.0695	0.0683	0.949
	(0.2, 0.8)	Oracle	-0.0064	0.0843	0.0837	0.935	0.0180	0.0663	0.0636	0.934
		Naive	-0.0070	0.0844	0.0837	0.939	0.1168	0.0453	0.0434	0.213
		Proposed	-0.0065	0.0842	0.0837	0.936	0.0160	0.0689	0.0664	0.935

“Oracle” denotes the oracle estimator calculated based on the true covariate. “Naive” denotes the naive estimator calculated based on the unadjusted covariate. “Proposed” denotes the proposed estimator calculated based on the adjusted covariate.

undertaken to see the effect of Type and FEV on the failure time as follows:

$$\lambda(t \mid \text{Type}, \text{FEV}) = \lambda_0(t) + \theta \text{Type} + \gamma \text{FEV}. \quad (69)$$

The Kaplan–Meier survival curves have been drawn with related theory taking the kinds of treatment types and the amount of FEV (adjusted FEV) of the patients into account. In the process of drawing, we view FEV (adjusted FEV) to be

1 when $\text{FEV (adjusted FEV)} \geq \text{the median of FEV (the median of adjusted FEV)}$, otherwise, to be 0. As shown in Figure 1, it can be seen that disturbance did affect relations between FEV and survival probability, and the patients with placebo or lower FEV (lower adjusted FEV) tend to have lower survival probabilities.

An analysis has been presented that it is available to derive the coefficients in model (69) with the proposed covariate-adjusted approach and summarize estimated

TABLE 3: Simulation results when $\lambda_0(t) = 3t^2 + 1$.

Cutpoints	Method	$\hat{\theta}$				$\hat{\gamma}$			
		Bias	SD	SE	CP	Bias	SD	SE	CP
$C \sim U[0, 1]I(H < 0) + U[0, 3]I(H \geq 0)$									
(0.3, 0.7)	Oracle	0.0015	0.1096	0.1146	0.955	0.0032	0.2418	0.2365	0.939
	Naive	0.0017	0.1100	0.1145	0.955	0.1220	0.1569	0.1509	0.865
	Proposed	0.0014	0.1096	0.1146	0.956	0.0080	0.2377	0.2319	0.934
(0.2, 0.8)	Oracle	0.0000	0.0927	0.0983	0.959	-0.0028	0.2050	0.2024	0.940
	Naive	0.0002	0.0927	0.0984	0.961	0.1188	0.1275	0.1288	0.865
	Proposed	0.0001	0.0928	0.0983	0.959	0.0046	0.2001	0.1982	0.942
$C \sim E(1)I(H < 0) + E(1/3)I(H \geq 0)$									
(0.3, 0.7)	Oracle	0.0044	0.1269	0.1265	0.950	-0.0078	0.2622	0.2604	0.942
	Naive	0.0046	0.1270	0.1268	0.949	0.1162	0.1691	0.1648	0.905
	Proposed	0.0042	0.1268	0.1265	0.949	0.0026	0.2556	0.2548	0.948
(0.2, 0.8)	Oracle	0.0051	0.1007	0.1064	0.957	-0.0156	0.2261	0.2170	0.944
	Naive	0.0046	0.1015	0.1065	0.955	0.1172	0.1394	0.1386	0.860
	Proposed	0.0050	0.1009	0.1064	0.957	-0.0077	0.2226	0.2126	0.943

“Oracle” denotes the oracle estimator calculated based on the true covariate. “Naive” denotes the naive estimator calculated based on the unadjusted covariate. “Proposed” denotes the proposed estimator calculated based on the adjusted covariate.

TABLE 4: Simulation results when $\lambda_0(t) = e^{-t} + t$.

Cutpoints	Method	$\hat{\theta}$				$\hat{\gamma}$			
		Bias	SD	SE	CP	Bias	SD	SE	CP
$C \sim U[0, 1]I(H < 0) + U[0, 3]I(H \geq 0)$									
(0.3, 0.7)	Oracle	0.0012	0.0810	0.0894	0.971	-0.0033	0.1812	0.1762	0.955
	Naive	0.0009	0.0804	0.0895	0.973	0.1154	0.1151	0.1112	0.843
	Proposed	0.0011	0.0809	0.0895	0.971	0.0061	0.1784	0.1726	0.956
(0.2, 0.8)	Oracle	-0.0045	0.0746	0.0773	0.955	-0.0002	0.1533	0.1540	0.949
	Naive	-0.0049	0.0747	0.0774	0.957	0.1154	0.0995	0.0973	0.791
	Proposed	-0.0045	0.0746	0.0773	0.957	0.0055	0.1523	0.1511	0.950
$C \sim E(1)I(H < 0) + E(1/3)I(H \geq 0)$									
(0.3, 0.7)	Oracle	-0.0001	0.0879	0.0870	0.943	0.0025	0.1836	0.1799	0.935
	Naive	-0.0002	0.0878	0.0872	0.943	0.1166	0.1126	0.1127	0.855
	Proposed	-0.0001	0.0877	0.0871	0.942	0.0099	0.1791	0.1762	0.941
(0.2, 0.8)	Oracle	0.0018	0.0738	0.0747	0.954	-0.0058	0.1598	0.1537	0.936
	Naive	0.0018	0.0743	0.0748	0.955	0.1180	0.0985	0.0960	0.771
	Proposed	0.0019	0.0739	0.0747	0.954	0.0013	0.1568	0.1505	0.936

“Oracle” denotes the oracle estimator calculated based on the true covariate. “Naive” denotes the naive estimator calculated based on the unadjusted covariate. “Proposed” denotes the proposed estimator calculated based on the adjusted covariate.

FIGURE 1: Estimated Kaplan–Meier survival curves for patients in pulmonary exacerbations study: (a) under different expiratory volumes; (b) under different adjusted expiratory volumes; (c) under different treatment types.

TABLE 5: Results for analysis of pulmonary exacerbation study data.

	Est	Naive Bias	SE	Est	Proposed Bias	SE
Cutpoints (0.3, 0.7)						
Type	-1.1342×10^{-5}	-1.8860×10^{-6}	6.1293×10^{-6}	-3.0592×10^{-5}	-1.3562×10^{-6}	1.0596×10^{-5}
FEV	-8.6200×10^{-4}	8.9298×10^{-5}	7.1152×10^{-4}	-8.5072×10^{-4}	-3.2173×10^{-6}	7.1324×10^{-4}
Cutpoints (0.2, 0.8)						
Type	-1.2598×10^{-5}	-6.2450×10^{-7}	6.0173×10^{-6}	-3.3581×10^{-5}	8.1670×10^{-7}	1.0456×10^{-5}
FEV	-9.5138×10^{-4}	9.9719×10^{-5}	6.9659×10^{-4}	-9.4777×10^{-4}	2.2955×10^{-6}	6.9800×10^{-4}

“Naive” denotes the naive estimator calculated based on the unadjusted covariate. “Proposed” denotes the proposed estimator calculated based on the adjusted covariate.

coefficients to be the column Est. After 1000 times artificially estimating the process, it shows the main characteristic of SE and the Bias which are calculated by the average of parameter estimates minus the corresponding Est. We also apply the contaminated covariate to calculate the estimator. The results based on our method are listed in the columns under Proposed and the results based on the contaminated covariate are put in the columns under Naive in Table 5. From the result above, we can see that, with the increase of the mount of FEV, the risk of relapse with pulmonary exacerbations will decrease. The treatment type rhDNase of pulmonary exacerbations can decrease the risk of death, which is consistent with Figure 1. Moreover, due to the covariate adjustment process, the bias of the proposed method is less than that of the Naive method. The sample standard error of the cutpoint (0.2, 0.8) is less than that of the cutpoint (0.3, 0.7), which is in accord with the simulation results.

6. Conclusion

From the point of view, the fact is that the ODS design is a benefit to lower the cost of expensive exposure variable and improve computing efficiency in large-scale cohort studies. All the results are performed for avoiding miscalculations and perfecting interpretation with adjustment on contaminated covariates. Firstly, this paper has illustrated the method of fitting covariate-adjusted additive hazard model to the data in the ODS design. Secondly, to solve the problems caused by contaminated covariates and biased-sampling schemes, this paper uses nonparametric kernel estimation method to calibrate the contaminated covariates and uses the inverse probability weighting method based on the calibration covariate to construct the weighted estimation function. In fact, this paper has carried out an extension of the theory properties of the estimator in the analysis of the proposed weighted estimation function. The numerical simulation studies have been applied to show that the estimator proposed in this paper performs well in finite sample case, and the actual data is used to show the possibility of the implement of the method.

From the perspective of research prospects, firstly, based on the methods and conclusions of this paper on the covariate adjustment problem of the additive hazard model, there has been an increasing awareness of the potential of discussing the problem of covariate adjustment of ODS

studies based on some other models. For example, the accelerated failure time model discussed by Lin et al. [29], the accelerated hazard model studied by Chen and Wang [30], the nonautonomous SIRS model discussed by Lv and Meng [31], and the dynamic model of the constantor studied in [32]. Secondly, to better promote the design and keep the proposed method practical, a promising research published by Yu et al. [16] who proposed that the determination of sample sizes and the optimal sample allocation method may also be an interesting topic in the future. Furthermore, it is hoped that future research will contribute to a further development of survival models with multiple disease outcomes mentioned by Kang and Cai [7].

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) (nos. 11901175 and 71901222), the Humanities and Social Science Foundation of the Ministry of Education of China (no. 17YJC630236), and Fundamental Research Funds for the Hubei Key Laboratory of Applied Mathematics, Hubei University (no. HBAM201907).

References

- [1] R. L. Prentice, “A case-cohort design for epidemiologic cohort studies and disease prevention trials,” *Biometrika*, vol. 73, no. 1, pp. 1–11, 1986.
- [2] S. G. Self and R. L. Prentice, “Asymptotic distribution theory and efficiency results for case-cohort studies,” *The Annals of Statistics*, vol. 16, no. 1, pp. 64–81, 1988.
- [3] W. Y. Tsai, “Pseudo-partial likelihood for proportional hazards models with biased-sampling data,” *Biometrika*, vol. 96, no. 3, pp. 601–615, 2009.
- [4] S. Kim, J. Cai, and W. Lu, “More efficient estimators for case-cohort studies,” *Biometrika*, vol. 100, no. 3, pp. 695–708, 2013.

- [5] H. Y. Chen, "Fitting semiparametric transformation regression models to data from a modified case-cohort design," *Biometrika*, vol. 88, no. 1, pp. 255–268, 2001.
- [6] J. Cai and D. Zeng, "Power calculation for case-cohort studies with nonrare events," *Biometrics*, vol. 63, no. 4, pp. 1288–1295, 2007.
- [7] S. Kang and J. Cai, "Marginal hazards model for case-cohort studies with multiple disease outcomes," *Biometrika*, vol. 96, no. 4, pp. 887–901, 2009.
- [8] N. Chatterjee, Y. H. Chen, and N. E. Breslow, "A pseudoscore estimator for regression problems with two-phase sampling," *Journal of the American Statistical Association*, vol. 98, no. 1, pp. 158–168, 2003.
- [9] H. Zhou, M. A. Weaver, J. Qin, M. P. Longnecker, and M. C. Wang, "A semiparametric empirical likelihood method for data from an outcome-dependent sampling scheme with a continuous outcome," *Biometrics*, vol. 58, no. 2, pp. 413–421, 2002.
- [10] N. A. Weaver and H. Zhou, "An estimated likelihood method for continuous outcome regression models with outcome-dependent sampling," *Journal of the American Statistical Association*, vol. 100, no. 470, pp. 459–469, 2005.
- [11] H. Zhou, G. Qin, and M. P. Longnecker, "A partial linear model in the outcome-dependent sampling setting to evaluate the effect of prenatal PCB exposure on cognitive function in children," *Biometrics*, vol. 67, no. 3, pp. 876–885, 2011.
- [12] H. Zhou, J. You, G. Qin, and M. P. Longnecker, "A partially linear regression model for data from an outcome-dependent sampling design," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 60, no. 4, pp. 559–574, 2011.
- [13] G. Qin and H. Zhou, "Partial linear inference for a 2-stage outcome-dependent sampling design with a continuous outcome," *Biostatistics*, vol. 12, no. 3, pp. 506–520, 2011.
- [14] S. Yan, J. Ding, and Y. Liu, "Statistical inference methods and applications of outcome-dependent sampling designs under generalized linear models," *Science China Mathematics*, vol. 60, no. 7, pp. 1219–1238, 2017.
- [15] J. H. Ding, L. Zhou, and J. Liu, "Estimating effect of environmental contaminants on women's subfecundity for the MoBa study data with an outcome-dependent sampling scheme," *Biostatistics*, vol. 15, no. 4, pp. 636–650, 2014.
- [16] J. Yu, Y. Liu, J. Cai, D. P. Sandler, and H. Zhou, "Outcome-dependent sampling design and inference for Cox's proportional hazards Model," *Journal of Statistical Planning and Inference*, vol. 178, pp. 24–36, 2016.
- [17] D. Şentürk and H. G. Müller, "Covariate adjusted correlation analysis via varying coefficient models," *Scandinavian Journal of Statistics*, vol. 32, no. 3, pp. 365–383, 2005.
- [18] D. Şentürk and H. G. Müller, "Covariate-adjusted regression," *Biometrika*, vol. 92, no. 1, pp. 75–89, 2005.
- [19] D. Şentürk and H. G. Müller, "Inference for covariate adjusted regression via varying coefficient models," *The Annals of Statistics*, vol. 34, no. 2, pp. 654–679, 2006.
- [20] X. Cui, W. Guo, L. Lin, and L. Zhu, "Covariate-adjusted nonlinear regression," *The Annals of Statistics*, vol. 37, no. 4, pp. 1839–1870, 2009.
- [21] F. Li, L. Lin, and X. Cui, "Covariate-adjusted partially linear regression models," *Communications in Statistics—Theory and Methods*, vol. 39, no. 6, pp. 1054–1074, 2010.
- [22] X. Cui, "Statistical analysis of two types of complex data and its associated model," Ph.D. thesis, Shandong University, Shandong, China, 2008.
- [23] J. Zhang, L.-P. Zhu, and L.-X. Zhu, "On a dimension reduction regression with covariate adjustment," *Journal of Multivariate Analysis*, vol. 104, no. 1, pp. 39–55, 2012.
- [24] A. Delaigle, P. Hall, and W.-X. Zhou, "Nonparametric covariate-adjusted regression," *The Annals of Statistics*, vol. 44, no. 5, pp. 2190–2220, 2016.
- [25] J. Zhao and C. Xie, "A nonparametric test for covariate-adjusted models," *Statistics & Probability Letters*, vol. 133, pp. 65–70, 2018.
- [26] D. Y. Lin and Z. Ying, "Semiparametric analysis of the additive risk model," *Biometrika*, vol. 81, no. 1, pp. 61–71, 1994.
- [27] J. Yu, Y. Liu, D. P. Sandler, and H. Zhou, "Statistical inference for the additive hazards model under outcome-dependent sampling," *Canadian Journal of Statistics*, vol. 43, no. 3, pp. 436–453, 2015.
- [28] P. K. Andersen and R. D. Gill, "Cox's regression model for counting processes: a large sample study," *The Annals of Statistics*, vol. 10, no. 4, pp. 1100–1120, 1982.
- [29] D. Lin, L. J. Wei, and Z. Ying, "Accelerated failure time models for counting processes," *Biometrika*, vol. 85, no. 3, pp. 605–618, 1998.
- [30] Y. Chen and M. Wang, "Analysis of accelerated hazards models," *Journal of the American Statistical Association*, vol. 95, no. 450, pp. 608–618, 2000.
- [31] X. Lv and X. Meng, "Analysis of dynamic behavior for a nonlinear stochastic non-autonomous SIRS model," *Mathematical Modeling and Its Applications*, vol. 7, no. 1, pp. 16–23+79, 2018.
- [32] T. Zhang, N. Gao, and J. Wang, "Dynamic system of microbial culture described by impulsive differential equations," *Mathematical Modeling and Its Applications*, vol. 8, no. 1, pp. 1–13, 2019.