Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEARCH ARTICLE

Sampling-Based Estimation for Massive Survival Data with Additive Hazards Model

Lulu Zuo[1] | Haixiang Zhang*[1] | HaiYing Wang[2] | Lei Liu[3]

[1]Center for Applied Mathematics, Tianjin University, Tianjin, 300072, China
[2]Department of Statistics, University of Connecticut, Storrs, Mansfield, CT 06269, USA
[3]Division of Biostatistics, Washington University in St. Louis, St. Louis, MO 63110, USA

Correspondence
*Haixiang Zhang, Center for Applied Mathematics, Tianjin University, Tianjin, 300072, China. Email: haixiang.zhang@tju.edu.cn

Summary

For massive survival data, we propose a subsampling algorithm to efficiently approximate the estimates of regression parameters in the additive hazards model. We establish consistency and asymptotic normality of the subsample-based estimator given the full data. The optimal subsampling probabilities are obtained via minimizing asymptotic variance of the resulting estimator. The subsample-based procedure can largely reduce the computational cost compared with the full data method. In numerical simulations, our method has low bias and satisfactory coverage probabilities. We provide an illustrative example on the survival analysis of patients with lymphoma cancer from the Surveillance, Epidemiology, and End Results Program.

KEYWORDS:
Additive hazards model; Big data; Subsample-based estimator; Subsampling probabilities; Survival analysis

1 | INTRODUCTION

Advancements in health information technology have led to an influx of massive data. One common feature of massive data is the huge number of observations (large n), which lays a heavy burden on storage and computation. In recent years substantial research effort has been devoted to the statistical analysis of massive data. For example, Zhao et al.[1] considered a partially linear framework for modeling massive heterogeneous data. Battey et al.[2] investigated hypothesis testing and parameter estimation using the “divide and conquer” algorithm. Shi et al.[3] studied the “divide and conquer" method for cubic-rate estimators. Jordan et al.[4] presented a communication-efficient surrogate likelihood method for distributed statistical inference problems. Volgushev et al.[5] proposed a two-step distributed inference for quantile regression with massive datasets.

Another approach to the analysis of massive data is subsampling, e.g. Ma et al.[6] proposed a leveraging-based subsampling procedure. Wang et al.[7] and Wang[8] developed optimal subsampling methods for logistic regression. Wang et al.[9] provided an information-based optimal subdata selection approach in the context of linear models. Wang and Ma[10] investigated optimal subsampling for quantile regression. Zhang and Wang[11] proposed a distributed subsampling procedure for big data linear models. Note that the “divide and conquer" method aims at analyzing the full data with parallel or distributed computing platforms, while the subsampling method focuses on fast calculation with limited computing resources in practical applications.

The above-mentioned articles are mainly focused on completely observed (uncensored) data. Only a limited number of papers have studied the topics on massive survival data. For example, Kawaguchi et al.[12] developed a new scalable sparse Cox regression method for high-dimensional survival data with massive sample sizes. Wang et al.[13] proposed an efficient “divide and conquer" algorithm to fit sparse Cox regression with massive datasets. Xue et al.[14] proposed an online updating approach for testing the proportional hazards assumption with streams of survival data.

As a competitive alternative to the Cox proportional hazards (PH) model, the additive hazards (AH) model (Aalen[15]; Lin and Ying[16]) has several advantages: examining additive associations vs. multiplicative associations, not assuming proportional hazards, and avoiding issues with the interpretation of the hazard ratio. These advantages may also scale well to the massive data case, while Xue et al.[14] demonstrated the complexity of examining PH with massive data. To the best of our knowledge, subsampling procedures have not been developed for censored survival data. In this paper, we propose a subsampling-based estimation method for massive survival data in the context of AH model. There are several advantages of our method. First, we propose a subsample-based estimator to approximate the full data estimator, and our method effectively reduces the computational CPU time. Second, the subsample-based estimator has an explicit expression, which is easy to calculate in practical applications. Third, we establish the asymptotic distribution of the subsample-based estimator given full data, which is very useful from the view of statistical inference.

The remainder of this paper is organized as follows. In Section 2, we review the AH model and propose a general subsampling algorithm. Asymptotic properties of the subsample estimator are established. In Section 3, we give a desirable subsampling strategy. In Section 4, we evaluate our method through numerical simulations. A real example of lymphoma cancer is illustrated in Section 5. Section 6 concludes this paper with some discussions. Technical proofs of theoretical results, Tables S.1 − S.6, an additional simulation study, and R codes for our proposed method are given in the *Supporting Information*.

2 | METHODS

2.1 | Notations and Estimation of AH Model

Let T_i be the failure time and C_i be the censoring time, $i = 1, \cdots, n$. Denote the observed follow-up time by $\tilde{T}_i = \min(T_i, C_i)$, where T_i and C_i are assumed to be independent in this paper. The failure indicator is $\Delta_i = I(T_i \leq C_i)$, and the censoring rate is $\delta = 1 - n^{-1} \sum_{i=1}^n \Delta_i$. Denote the observed-failure counting process by $N_i(t) = I(\tilde{T}_i \leq t, \Delta_i = 1)$, and the at-risk indicator by $Y_i(t) = I(\tilde{T}_i \geq t)$. Following Lin and Ying,[16] the intensity of $N_i(t)$ with additive hazards function is

$$d\Lambda_i(t) = Y_i(t)\{d\Lambda_0(t) + \boldsymbol{\theta}'\mathbf{X}_i dt\}, 1 \leq i \leq n, \tag{1}$$

where $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)'$ is a vector of regression parameters belonging to a compact subset of $\mathbb{R}^p$, $\mathbf{X}_i = (X_{i1}, \cdots, X_{ip})'$ is a vector of covariates, and $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$ is an unknown baseline cumulative hazards function. From Lin and Ying,[16] an estimator $\hat{\boldsymbol{\theta}}_{ZE}$ can be obtained by solving the estimating equation $\Psi(\boldsymbol{\theta}) = 0$, where

$$\Psi(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \int_0^\tau \{\mathbf{X}_i - \bar{\mathbf{X}}(t)\}\{dN_i(t) - Y_i(t)\boldsymbol{\theta}'\mathbf{X}_i dt\}. \tag{2}$$

Here $\bar{\mathbf{X}}(t) = \sum_{i=1}^n Y_i(t)\mathbf{X}_i / \sum_{i=1}^n Y_i(t)$, and $\tau > 0$ is the length of the study. For convenience, denote the full data by $\mathcal{F}_n = (\mathbf{X}_{full}, \tilde{\mathbf{T}}_{full}, \boldsymbol{\Delta}_{full})$, where $\mathbf{X}_{full} = (\mathbf{X}_1, \cdots, \mathbf{X}_n)'$ is the covariate matrix, $\tilde{\mathbf{T}}_{full} = (\tilde{T}_1, \cdots, \tilde{T}_n)$ consists of the observed follow-up times, and $\boldsymbol{\Delta}_{full} = (\Delta_1, \cdots, \Delta_n)$ consists of the failure indicators. Furthermore, $(\mathbf{X}_i, \tilde{T}_i, \Delta_i)$ are independent observations, $i = 1, \cdots, n$. We rewrite (2) as

$$\Psi(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \psi_i(\boldsymbol{\theta}), \tag{3}$$

where $\psi_i(\boldsymbol{\theta}) = \int_0^\tau \{\mathbf{X}_i - \bar{\mathbf{X}}(t)\}\{dN_i(t) - Y_i(t)\boldsymbol{\theta}'\mathbf{X}_i dt\}$, $i = 1, \cdots, n$. When the sample size n is very large, it is time-consuming to calculate $\hat{\boldsymbol{\theta}}_{ZE}$ due to the heavy computational burden. To deal with this problem, we propose a subsampling-based procedure. The basic idea is as follows: assign subsampling probabilities $\pi_i > 0$ for full data $(\mathbf{X}_i, \tilde{T}_i, \Delta_i)$ with $\sum_{i \in S_0} \pi_i = \delta$ and $\sum_{i \in S_1} \pi_i = 1 - \delta$, where δ is the censoring rate, $S_0 = \{i : \Delta_i = 0\}$ and $S_1 = \{i : \Delta_i = 1\}$ are the index sets of censored and noncensored individuals, respectively. Draw a random subsample of size $r (\ll n)$ from the full data with replacement according to subsampling probabilities $\{\pi_i\}_{i=1}^n$. Denote the corresponding subsample as $(\mathbf{X}_i^*, \tilde{T}_i^*, \Delta_i^*)$ with subsampling probabilities π_i^*, for $i = 1, \cdots, r$. Based on this subsample, we propose a weighted estimating function

$$\mathbf{U}^*(\boldsymbol{\theta}) = \frac{1}{nr}\sum_{i=1}^r \frac{1}{\pi_i^*} U_i^*(\boldsymbol{\theta}), \tag{4}$$

where $U_i^*(\boldsymbol{\theta}) = \int_0^\tau \{\mathbf{X}_i^* - \bar{\mathbf{X}}^*(t)\}\{dN_i^*(t) - Y_i^*(t)\boldsymbol{\theta}'\mathbf{X}_i^* dt\}$, with $\bar{\mathbf{X}}^*(t) = \{\sum_{i=1}^r \pi_i^{*-1} Y_i^*(t)\mathbf{X}_i^*\}/\{\sum_{i=1}^r \pi_i^{*-1} Y_i^*(t)\}$, $N_i^*(t) = I(\tilde{T}_i^* \leq t, \Delta_i^* = 1)$ and $Y_i^*(t) = I(\tilde{T}_i^* \geq t)$, $i = 1, \cdots, r$. Later we will show that $\mathbf{U}^*(\boldsymbol{\theta})$ is asymptotically unbiased towards

(3) given $\mathcal{F}_n$. Hence, we can get a subsample-based estimator $\tilde{\boldsymbol{\theta}}$ by solving $\mathbf{U}^*(\boldsymbol{\theta}) = 0$, and use $\tilde{\boldsymbol{\theta}}$ to approximate the full data estimate $\hat{\boldsymbol{\theta}}_{ZE}$. Our method can effectively reduce the computational burden, and the comparison of CPU time is given in the simulation section.

2.2 | Subsampling Algorithm and Asymptotic Properties

In this section, we propose a subsampling algorithm for the subsample estimator $\tilde{\boldsymbol{\theta}}$ as follows:

Algorithm 1 Subsampling Algorithm

Step 1 (*Sampling*): assign subsampling probabilities $\pi_i > 0$ for the full data $\mathcal{F}_n$ with $\sum_{i \in S_0} \pi_i = \delta$ and $\sum_{i \in S_1} \pi_i = 1 - \delta$. Draw a random subsample of size $r(\ll n)$ from the full data with replacement according to $\{\pi_i\}_{i=1}^n$. Denote the corresponding subsample as $(\mathbf{X}_i^*, \tilde{T}_i^*, \Delta_i^*)$ together with π_i^*, for $i = 1, \cdots, r$.
Step 2 (*Estimation*): We obtain a subsampling-based estimator $\tilde{\boldsymbol{\theta}}$ satisfying $\mathbf{U}^*(\tilde{\boldsymbol{\theta}}) = 0$ with the subsample in Step 1, where $\tilde{\boldsymbol{\theta}}$ has an explicit expression

$$\tilde{\boldsymbol{\theta}} = \left[\frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}\int_0^\tau Y_i^*(t)\{\mathbf{X}_i^* - \bar{\mathbf{X}}^*(t)\}^{\otimes 2} dt\right]^{-1}\left[\frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}\int_0^\tau \{\mathbf{X}_i^* - \bar{\mathbf{X}}^*(t)\} dN_i^*(t)\right], \tag{5}$$

where $\mathbf{c}^{\otimes 2} = \mathbf{c}\mathbf{c}'$ for a vector $\mathbf{c}$.

Given $\mathcal{F}_n$, the consistency and asymptotic normality of $\tilde{\boldsymbol{\theta}}$ are needed to determine the optimal subsampling probabilities in Section 3. Under assumptions (A.1)–(A.7) in the *Supporting Information*, as $n \to \infty$ and $r \to \infty$, for any $\epsilon > 0$, with probability approaching one, there exist finite Δ_ϵ and r_ϵ, such that

$$P(\|\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{ZE}\| \geq r^{-1/2}\Delta_\epsilon | \mathcal{F}_n) < \epsilon, \tag{6}$$

for all $r \geq r_\epsilon$. This consistency ensures that we can efficiently approximate $\hat{\boldsymbol{\theta}}_{ZE}$ by the subsample-based estimator $\tilde{\boldsymbol{\theta}}$. Hence, we use $\tilde{\boldsymbol{\theta}}$ rather than $\hat{\boldsymbol{\theta}}_{ZE}$ to reduce the computational burden.

Next, we establish the asymptotic normality of $\tilde{\boldsymbol{\theta}}$. Under assumptions (A.1)–(A.8) in the *Supporting Information*, as $n \to \infty$ and $r \to \infty$, conditional on $\mathcal{F}_n$, we have

$$\boldsymbol{\Sigma}^{-1/2}(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{ZE}) \xrightarrow{d} N(0, \mathbf{I}), \tag{7}$$

where $\xrightarrow{d}$ denotes convergence in distribution, $\boldsymbol{\Sigma} = \mathcal{H}^{-1}\boldsymbol{\Gamma}\mathcal{H}^{-1}$ with

$$\mathcal{H} = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau Y_i(t)\{\mathbf{X}_i - \bar{\mathbf{X}}(t)\}^{\otimes 2} dt, \tag{8}$$

and

$$\boldsymbol{\Gamma} = \frac{1}{rn^2}\sum_{i=1}^{n}\frac{1}{\pi_i}\int_0^\tau \{\mathbf{X}_i - \bar{\mathbf{X}}(t)\}^{\otimes 2} dN_i(t). \tag{9}$$

3 | SUBSAMPLING STRATEGIES

We consider how to specify the subsampling probablities $\{\pi_i\}_{i=1}^n$. A naive choice is the uniform subsampling strategy with $\pi_i = n^{-1}$, for $i = 1, \cdots, n$. However, these uniform subsampling probabilities may not be optimal, and a nonuniform subsampling method could have a better performance.[7] Our idea is to determine the optimal subsampling probabilities by minimizing the asymptotic variance matrix $\boldsymbol{\Sigma}$ of $\tilde{\boldsymbol{\theta}}$ in (7). Since $\boldsymbol{\Sigma}$ is a matrix, the meaning of "minimizing" needs to be carefully defined. For this purpose, we use the trace to induce a complete ordering of the asymptotic variance matrix.[17] The asymptotic mean squared error (AMSE) of $\tilde{\boldsymbol{\theta}}$ is equal to the trace of $\boldsymbol{\Sigma}$, which is given by

$$\mathrm{AMSE}(\tilde{\boldsymbol{\theta}}) = tr(\boldsymbol{\Sigma}), \tag{10}$$

where $tr(\cdot)$ denotes the trace of a matrix.

As mentioned above, the subsampling probabilities derived by minimizing $tr(\boldsymbol{\Sigma})$ require the calculation of $\mathcal{H}^{-1}$, which takes substantial time in the case of large n. Because $\mathcal{H}$ and $\boldsymbol{\Gamma}$ are nonnegative definite, and $\boldsymbol{\Sigma} = \mathcal{H}^{-1}\boldsymbol{\Gamma}\mathcal{H}^{-1}$, simple matrix algebra yields that $tr(\boldsymbol{\Sigma}) = tr(\boldsymbol{\Gamma}\mathcal{H}^{-2}) \leq [tr(\boldsymbol{\Gamma}^2)]^{1/2}[tr(\mathcal{H}^{-4})]^{1/2} \leq tr(\boldsymbol{\Gamma})tr(\mathcal{H}^{-2}) \leq n\lambda_{max}(\mathcal{H}^{-2})tr(\boldsymbol{\Gamma})$, where $\lambda_{max}(\cdot)$ denotes the maximum eigenvalue of a matrix. That is, the minimizer of $tr(\boldsymbol{\Gamma})$ minimizes an upper bound of $tr(\boldsymbol{\Sigma})$. In fact, $\boldsymbol{\Sigma}$ depends on $\boldsymbol{\pi}_i$ only through $\boldsymbol{\Gamma}$, and $\mathcal{H}$ is free of $\boldsymbol{\pi}_i$. Hence, we suggest to determine the subsampling probabilites by directly minimizing $tr(\boldsymbol{\Gamma})$, which can effectively speed up the subsampling algorithm. Note that

$$\begin{aligned}
tr(\boldsymbol{\Gamma}) &= tr\left(\frac{1}{rn^2}\sum_{i=1}^{n}\frac{\int_0^\tau\{\mathbf{X}_i-\bar{\mathbf{X}}(t)\}^{\otimes 2}dN_i(t)}{\pi_i}\right)\\
&= \frac{1}{rn^2}\sum_{i=1}^{n}\frac{tr(\int_0^\tau\{\mathbf{X}_i-\bar{\mathbf{X}}(t)\}^{\otimes 2}dN_i(t))}{\pi_i}\\
&= \frac{1}{rn^2}\left[\sum_{i\in S_0}\frac{tr(\int_0^\tau\{\mathbf{X}_i-\bar{\mathbf{X}}(t)\}^{\otimes 2}dN_i(t))}{\pi_i}+\sum_{i\in S_1}\frac{tr(\int_0^\tau\{\mathbf{X}_i-\bar{\mathbf{X}}(t)\}^{\otimes 2}dN_i(t))}{\pi_i}\right]\\
&= \frac{1}{rn^2}\sum_{i\in S_1}\frac{tr(\int_0^\tau\{\mathbf{X}_i-\bar{\mathbf{X}}(t)\}^{\otimes 2}dN_i(t))}{\pi_i}.
\end{aligned}$$

Due to $dN_i(t) = 0$ for $i \in S_0$, the corresponding subsampling probabilities $\{\pi_i\}_{i\in S_0}$ are not included in $tr(\boldsymbol{\Gamma})$. Hence, we cannot determine $\{\pi_i\}_{i\in S_0}$ by minimizing $tr(\boldsymbol{\Gamma})$. We point out that $\pi_i > 0$ is a basic requirement to ensure the asymptotic unbiasedness of $\mathbf{U}^*(\boldsymbol{\theta})$. In this case, one choice for the subsampling probabilities of censored individuals is $\pi_i^{m\Gamma} = \delta/K$ for $i \in S_0$, where K denotes the number of elements in S_0. Till now, the key point is to assign subsampling probabilities for non-censored individuals. The following result gives the subsampling probabilities $\pi_i^{m\Gamma}$ for $i \in S_1$.

Under assumptions (A.1)−(A.8) in the *Supporting Information*, if the subsampling probabilities are chosen as

$$\pi_i^{m\Gamma} = (1-\delta)\cdot\frac{tr^{1/2}\{\int_0^\tau\{\mathbf{X}_i-\bar{\mathbf{X}}(t)\}^{\otimes 2}dN_i(t)\}}{\sum_{i\in S_1}tr^{1/2}\{\int_0^\tau\{\mathbf{X}_i-\bar{\mathbf{X}}(t)\}^{\otimes 2}dN_i(t)\}},\quad \text{for } i\in S_1 \tag{11}$$

then $tr(\boldsymbol{\Gamma})$ attains its minimum, where $\delta = 1-n^{-1}\sum_{i=1}^n\Delta_i$ is the censoring rate. Of note, since $\sum_{i\in S_0}\pi_i = \delta$ and $\sum_{i\in S_1}\pi_i = 1-\delta$, a subsample has a similar censoring rate with the full data. In this case, a subsample can potentially capture the censoring property of the full data. Numerical simulation indicates that this choice works well in practice.

In what follows, the subsample estimator $\breve{\boldsymbol{\theta}}$ can be obtained by replacing π_i with $\pi_i^{m\Gamma}$ in (5), $i = 1,\cdots,n$. To reduce the computational burden, we propose to estimate the covariance matrix of $\breve{\boldsymbol{\theta}}$ with one subsample as follows:

$$\breve{\boldsymbol{\Sigma}} = \breve{\mathcal{H}}^{-1}\breve{\boldsymbol{\Gamma}}\breve{\mathcal{H}}^{-1}, \tag{12}$$

where

$$\breve{\mathcal{H}} = \frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}\int_0^\tau Y_i^*(t)\{\mathbf{X}_i^*-\bar{\mathbf{X}}^*(t)\}^{\otimes 2}dt,$$

$$\breve{\boldsymbol{\Gamma}} = \frac{1}{n^2r^2}\sum_{i=1}^{r}\frac{1}{\pi_i^{*2}}\int_0^\tau\{\mathbf{X}_i^*-\bar{\mathbf{X}}^*(t)\}^{\otimes 2}dN_i^*(t),$$

and $\{\pi_i^*\}_{i=1}^r$ are the corresponding subsampling probabilities for a subsample. The standard errors of components in $\breve{\boldsymbol{\theta}}$ are the square roots of the diagonal elements of $\breve{\boldsymbol{\Sigma}}$. We will evaluate the performance of (12) using numerical simulations in Section 4.

4 | NUMERICAL STUDIES

In this section, we conduct three simulation studies to assess 1) our method's performance with optimal and uniform subsampling probabilities in comparison to the full data approach, 2) the gain in computation time, and 3) our method's performance with mild

Table 1. Simulation results on the subsample estimator $\breve{\boldsymbol{\theta}}$ with Case I‡.

		OSP				UNIF			
	r	bias	ESE	SSE	CP	bias	ESE	SSE	CP
$\theta_1 = -1$	100	0.0465	0.2565	0.2483	0.961	0.0642	0.2665	0.2656	0.952
	300	0.0177	0.1378	0.1339	0.963	0.0184	0.1426	0.1423	0.939
	500	0.0101	0.1054	0.1101	0.940	0.0136	0.1087	0.1155	0.945
$\theta_2 = -0.5$	100	0.0273	0.2146	0.2139	0.954	0.0322	0.2234	0.2303	0.956
	300	0.0074	0.1126	0.1135	0.955	0.0100	0.1155	0.1080	0.966
	500	0.0035	0.0852	0.0849	0.960	0.0034	0.0875	0.0889	0.951
$\theta_3 = 0$	100	0.0001	0.1908	0.1871	0.959	0.0043	0.1957	0.2026	0.945
	300	0.0029	0.0984	0.0975	0.945	0.0030	0.1002	0.1022	0.948
	500	0.0006	0.0744	0.0723	0.959	0.0006	0.0760	0.0761	0.946
$\theta_4 = 0.5$	100	0.0238	0.2120	0.2054	0.965	0.0333	0.2186	0.2174	0.957
	300	0.0126	0.1115	0.1132	0.952	0.0176	0.1146	0.1192	0.938
	500	0.0079	0.0846	0.0883	0.939	0.0078	0.0870	0.0890	0.961
$\theta_5 = 1$	100	0.0519	0.2547	0.2466	0.966	0.0613	0.2670	0.2742	0.957
	300	0.0236	0.1376	0.1364	0.951	0.0290	0.1420	0.1462	0.943
	500	0.0124	0.1047	0.1055	0.946	0.0128	0.1088	0.1123	0.937

‡ "OSP" denotes the proposed method with optimal subsampling probabilities; "UNIF" denotes the proposed method with uniform subsampling probabilities; "bias" denotes the sample mean of the estimates minus the estimator $\hat{\theta}_{ZE}$; "ESE" denotes the estimated standard error of the estimates; "SSE" denotes the sampling standard error of the estimates; "CP" denotes the empirical 95% coverage probability towards $\hat{\theta}_{ZE}$.

vs. heavy censoring and how the censoring proportion could affect the choice of r. First, we generate failure times $(T_1, \cdots, T_n)$ from the AH model with hazards function $\lambda(t|\mathbf{X}) = 1 + \boldsymbol{\theta}'\mathbf{X}$, where the true parameter is $\boldsymbol{\theta} = (-1, -0.5, 0, 0.5, 1)^T$ with $p = 5$. We consider four cases for the generation of covariate matrix $\mathbf{X}$,

Case I : $\mathbf{X} \sim N(0, \boldsymbol{\Sigma})$, where $\Sigma_{ij} = 0.5^{|i-j|}$.

Case II: $\mathbf{X} \sim N(0, \boldsymbol{\Sigma})$, where $\Sigma_{ij} = 0.5^{I(i \neq j)}$.

Case III: $\mathbf{X} = (X_1, \cdots, X_5)^T$, and X_i are independent exponential random variables with probability density function $f(x) = 2e^{-2x}I(x > 0)$, $i = 1, \cdots, 5$.

Case IV: $\mathbf{X} \sim t_5(0, \boldsymbol{\Sigma})$, where $\mathbf{X}$ follows a multivariate t distribution with degree 5 and covariance matrix $\Sigma_{ij} = 0.5^{|i-j|}$.

Note that the above Cases I and II are symmetric, Case III is asymmetric, and Case IV is heavy-tailed. The censoring time C_i are generated from the uniform distribution over $(0, 3)$, which leads to about 28% censoring rate. The observed follow-up times are $\tilde{T}_i = \min(T_i, C_i)$, for $i = 1, \cdots, n$. We carry out computation on a sever with 128GB memory using R software. In Table 1, we report the estimation results from "the proposed method with optimal subsampling probabilities (OSP)" vs. "the proposed method with uniform subsampling probabilities (UNIF)" for Case I (other cases are given in Tables S.1 − S.3 of the *Supporting Information*) including the estimated bias (bias) given by the sample mean of the estimates minus the full data estimator $\hat{\boldsymbol{\theta}}_{ZE}$, the estimated standard error (ESE) of the estimates, the sampling standard error (SSE) of the estimates, and the empirical 95% coverage probability (CP). Given $\mathcal{F}_n$, the above simulation results are based on L =1000 replications with $n = 10^5$, $r = 100$, 300 and 500. It can be seen from the results that both estimators are unbiased. The ESE and SSE of subsample estimator are close to each other, and the coverage probabilities are satisfactory. Their performances become better as the subsample size r increases. Moreover, both ESE and SSE of the OSP-based estimates are smaller than those of UNIF-based method.

For further comparison, let

$$\text{MSE} = \frac{1}{L}\sum_{\ell=1}^{L} \|\breve{\boldsymbol{\theta}}^{(\ell)} - \hat{\boldsymbol{\theta}}_{ZE}\|^2, \tag{13}$$

where $\breve{\boldsymbol{\theta}}^{(\ell)}$ is from the ℓth replication, $\ell = 1, \cdots, L$. In Figure 1, we present the MSEs of each method. From the results, we can see that the MSEs of OSP are smaller than those of UNIF. To evaluate the estimation performances of OSP and UNIF towards

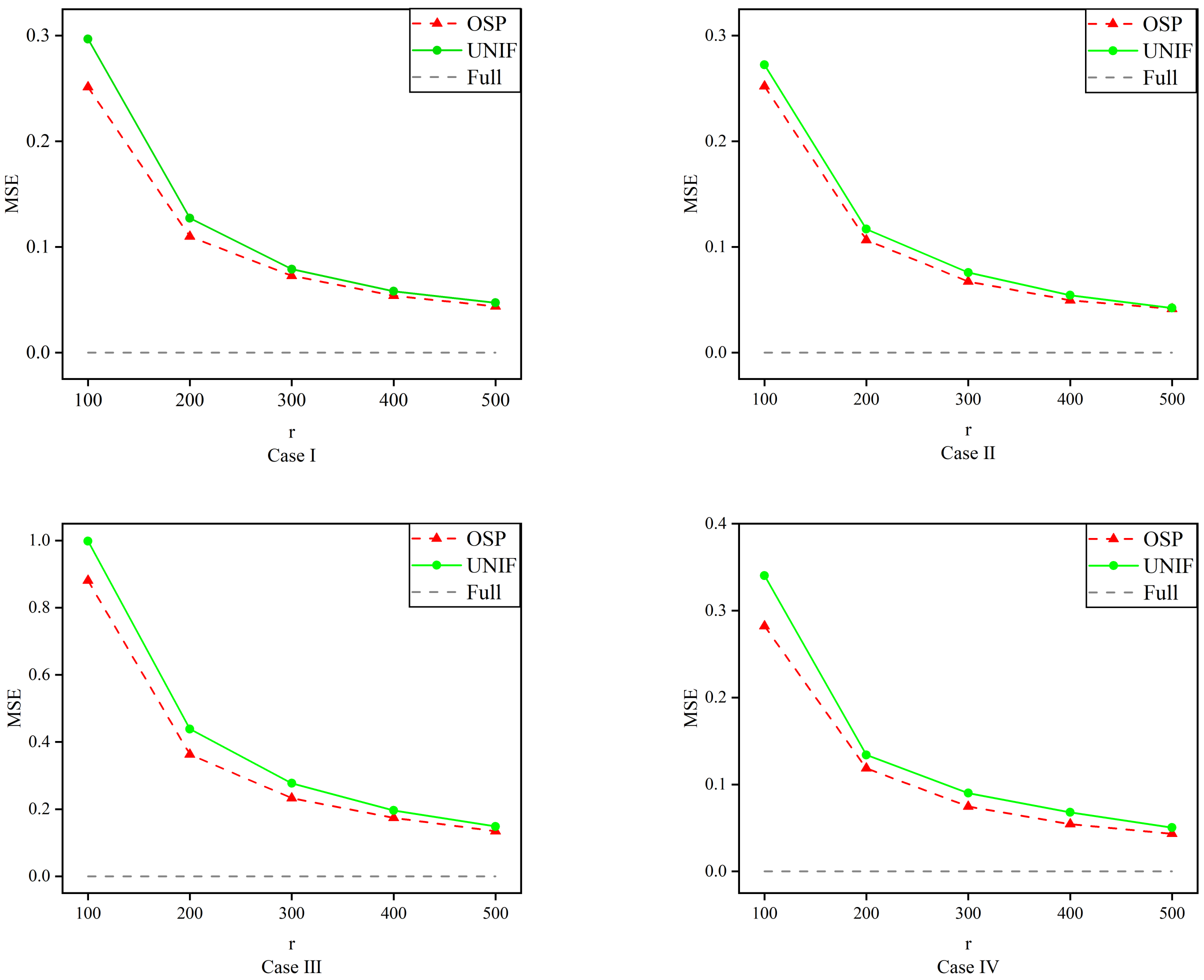

Figure 1. The MSEs for different subsampling methods.

different distribution of covariates, we define the estimation efficiency of OSP-based estimator relative to UNIF as

$$\text{Relative } Efficiency = \frac{\text{MSE}(\breve{\theta}_{unif})}{\text{MSE}(\breve{\theta}_{osp})},$$

where MSE is define in (13), $\breve{\theta}_{unif}$ and $\breve{\theta}_{osp}$ are the subsample estimators with UNIF and OSP, respectively. Figure 2 presents the relative efficiency towards different settings of covariates. We can conclude that $\breve{\theta}_{osp}$ is more efficient than $\breve{\theta}_{unif}$, especially in Cases III and IV.

Table 2. The CPU time for Case I with $r = 100$ (seconds)[‡].

	n			
Method	10^4	2×10^4	5×10^4	10^5
UNIF	13.847	13.870	13.896	13.926
OSP	21.990	26.349	51.674	148.560
Full data	40.853	115.781	871.220	4476.960

[‡] "OSP" and "UNIF" are given in the footnotes of Table 1.

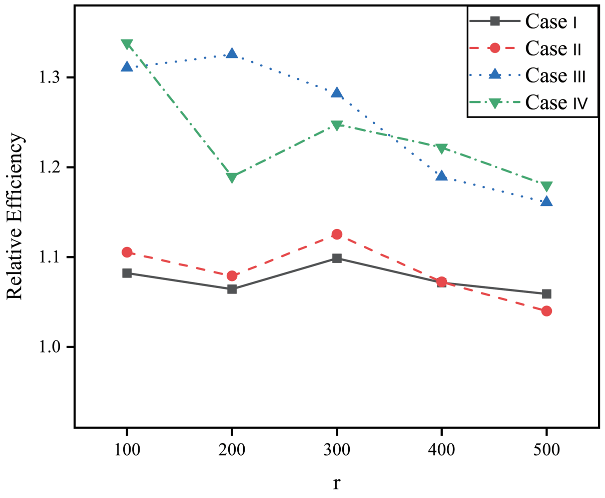

Figure 2. Relative efficiency for different settings of covariates.

We conduct the second simulation to evaluate the computational efficiency of the proposed subsampling algorithm, where the mechanism of data generation is the same as the first simulation. For fair comparison, we record the CPU time with one core based on the mean calculation time of 1000 repetitions of each subsample-based method. In Table 2, we report the results for the computing time for Case I with $r = 100$, $n = 10^4$, 2×10^4, 5×10^4 and 10^5. The computing time for the full data method is given in the last row. The UNIF requires the least computing time, because its subampling probabilities, $\pi_i = 1/n$, do not take time to compute. Note that the computational burden for the full data method is heavy, e.g. the CPU time is about 4476 seconds ($n = 10^5$). As the sample size n increases, the computational advantage of our proposed method becomes more convincing. Moreover, in Table 3 we report the computing time for Case I with $n = 10^5$, r=200, 400, 600, 800 and 1000, respectively. The results also indicate that our subsampling-based algorithm has great computation advantages over the full data method.

Table 3. The CPU time for Case I with $n = 10^5$ (seconds)[‡].

	r				
Method	200	400	600	800	1000
UNIF	14.012	14.419	14.617	14.903	15.363
OSP	149.952	150.439	152.584	153.621	155.384
Full data	4476.960				

[‡] “OSP" and “UNIF" are given in the footnotes of Table 1.

We conduct the third simulation to evaluate how the subsample-based method performs with different censoring rates. The simulation settings are the same as the first simulation, except that censoring times are generated from uniform distributions over (0, 6), (0, 3) and (0, 2), with corresponding censoring rate 16%, 28% and 38%, respectively. In Table 4, we report the bias, ESE, SSE and CP of the OSP-based subsample estimate $\breve{\theta}_1$ with Case I (other cases are given in Tables S.4 − S.6 of the *Supporting Information*), where $\breve{\theta}_i$ are similar and omitted, for $i = 2, \cdots, 5$. It can be seen from the results that the ESE and SSE become larger as the censoring rate δ increases. Hence, we suggest to use a larger subsample size r if the survival data is heavily censored in practice.

Table 4. Simulation results on OSP-based $\breve{\theta}_1$ under varying censoring rates (Case I)[‡].

	δ	bias	ESE	SSE	CP
$r = 100$	16%	0.0468	0.2393	0.2427	0.951
	28%	0.0465	0.2565	0.2483	0.961
	38%	0.0486	0.2917	0.2965	0.946
$r = 300$	16%	0.0089	0.1282	0.1238	0.953
	28%	0.0177	0.1378	0.1339	0.963
	38%	0.0259	0.1586	0.1664	0.935
$r = 500$	16%	0.0099	0.0980	0.0913	0.958
	28%	0.0101	0.1054	0.1101	0.940
	38%	0.0011	0.1205	0.1189	0.954

‡ δ is the censoring rate; "Bias", "ESE", "SSE" and "CP" are given in the footnotes of Table 1.

5 | A REAL DATA EXAMPLE

We apply our proposed method to a lymphoma cancer dataset in the Surveillance, Epidemiology, and End Results program (*https://seer.cancer.gov/*). There were 111,283 lymphoma cancer patients with full information between 1975 to 2007 in USA. For analysis, we set the censoring time as the first 60 months after being diagnosed as lymphoma cancer. Among those 111,283 subjects, the total number of event is 46,067 and the censoring rate is 58.6%. The risk factors $X_i = (X_{i1}, X_{i2})'$ are age (centered and scaled) and biological sex (male=1 and female=0). Our task is to approximate the $\hat{\boldsymbol{\theta}}_{ZE}$ in model (1) with our subsample-based method.

For comparison, we also report the full data based estimate $\hat{\boldsymbol{\theta}}_{ZE} = (\hat{\theta}_1, \hat{\theta}_2)'$ with $\hat{\theta}_1 = 0.0077$ and $\hat{\theta}_2 = 0.0011$, respectively. In Table 5, we report the the subsample estimator (Est), the standard error (SE) and the 95% confidence interval towards $\hat{\boldsymbol{\theta}}_{ZE}$ (CI) with one subsample, where the subsample size $r = 200$, 400 and 600, respectively. The results in Table 5 indicate that both UNIF and OSP based estimators are close to $\hat{\boldsymbol{\theta}}_{ZE}$. The SEs of OSP-based estimators are smaller than those of UNIF. The effects of age and gender are positive, which agree with the findings in Mukhtar et al.[18] Moreover, it seems that age (θ_1) is a significant risk factor. To further check the rationality of our method, we give bias, ESE and SSE of the subsample-based estimates based on 1000 subsamples in Table 6, where $r = 200$, 400 and 600, respectively. It can be seen from the results that both subsample-based estimators are unbiased, and the ESE is close to SSE. Hence, it is desirable to use one subsample with our method when analyzing real data in practice.

Table 5. Estimation results for the lymphoma cancer data with one subsample[‡].

		UNIF			OSP		
	θ	Est	SE	CI	Est	SE	CI
$r = 200$	θ_1	0.0065	0.0011	(0.0045, 0.0099)	0.0079	0.0010	(0.0060, 0.0085)
	θ_2	0.0005	0.0023	(−0.0033, 0.0067)	0.0017	0.0019	(−0.0006, 0.0042)
$r = 400$	θ_1	0.0077	0.0009	(0.0059, 0.0096)	0.0079	0.0008	(0.0062, 0.0095)
	θ_2	0.0017	0.0017	(−0.0016, 0.0050)	0.0011	0.0016	(−0.0021, 0.0043)
$r = 600$	θ_1	0.0081	0.0007	(0.0068, 0.0095)	0.0075	0.0006	(0.0062, 0.0085)
	θ_2	0.0002	0.0014	(−0.0022, 0.0026)	0.0011	0.0012	(−0.0035, 0.0019)

[‡] Est: the subsample estimator; SE: the standard error; CI: the 95% confidence interval towards $\hat{\theta}_{ZE}$.

6 | CONCLUDING REMARKS

In this paper, we have proposed a subsampling algorithm for the AH model with massive survival data. The subsample-based method can effectively approximate the full data estimator. The main advantage of our method is its much reduced computational

Table 6. Bias and (ESE, SSE) for the lymphoma cancer data[‡].

	θ	UNIF	OSP
$r = 200$	θ_1	−0.00016 (0.00123, 0.00125)	−0.00009 (0.00113, 0.00122)
	θ_2	−0.00014 (0.00239, 0.00242)	−0.00011 (0.00222, 0.00233)
$r = 400$	θ_1	−0.00018 (0.00122, 0.00122)	−0.00009 (0.00112, 0.00118)
	θ_2	0.00012 (0.00238, 0.00237)	−0.000004 (0.00222, 0.00232)
$r = 600$	θ_1	−0.00003 (0.00070, 0.00071)	−0.00002 (0.00064, 0.00068)
	θ_2	0.00002 (0.00136, 0.00145)	0.00001 (0.00127, 0.00135)

[‡] "Bias", "ESE", "SSE", "UNIF" and "OSP" are given in the footnotes of Table 1.

burden. From the view of statistical efficiency, the OSP-based estimator has a smaller SE than the UNIF method. Hence, we recommend the OSP when applying our method in practical applications. In conclusion, it is desirable to choose our subsampling approach over the methods of Kawaguchi et al.[12] or Xue et al.[14] when we have limited computing resources at hand.

Of note, the UNIF approach is different from bootstrap. Specifically, the UNIF method uses one subsample to approximate the full data estimator, and its main purpose is to reduce the computational time. However, the classic bootstrap needs many samples with full-size by repeatedly sampling, which aims to conduct statistical inference (e.g. estimating standard errors or confidence intervals). To further improve our method, we can consider an iterative subsampling procedure. Specifically, we perform L replications of our proposed approach. Let $\tilde{\boldsymbol{\theta}} = \frac{1}{L}\sum_{\ell=1}^{L} \breve{\boldsymbol{\theta}}^{(\ell)}$, where $\breve{\boldsymbol{\theta}}^{(\ell)}$ is the subsampling-based estimator from the ℓth replication, for $\ell = 1, \cdots, L$. The asymptotic properties of $\tilde{\boldsymbol{\theta}}$ needs further research. Second, the simulations and real data example indicate that the proposed method works well with a moderate subsample size (e.g., $r = 500$). Our method has a higher estimation efficiency with a larger subsample, while it requires more computing resource. Hence, the recommended subsample size is taken according to the available computing resource at hand. Third, it is interesting to extend our proposed methods to other survival models, such as the Cox model[19] and the accelerated failure time model.[20] Fourth, a known limitation of the additive hazards approach is that the hazard is not constrained to be positive. Therefore, it is interesting to assess the model fit or appropriateness of the additive hazards model in the massive data setting.

ACKNOWLEDGMENTS

The authors would like to thank the Editor, the Associate Editor and two reviewers for their constructive and insightful comments that greatly improved the manuscript. We also thank the SEER Program for permitting our access to the lymphoma cancer data. The work of Wang was supported by National Science Foundation (NSF), USA grant DMS-1812013. The work of Liu was supported by NIH UL1 TR002345. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NSF and NIH.

SUPPORTING INFORMATION

The *Supporting Information* contains technical proofs for the theoretical results, Tables S.1 − S.6 for the main document of our article, an additional simulation study, together with an R function `bigAH.R` for implementing our proposed method.

References

1. Zhao T, Cheng G, Liu H. A partially linear framework for massive heterogeneous data. *The Annals of Statistics* 2016; 44(4): 1400–1437.

2. Battey H, Fan J, Liu H, Lu J, Zhu Z. Distributed testing and estimation under sparse high dimensional models. *The Annals of Statistics* 2018; 46(3): 1352–1382.

3. Shi C, Lu W, Song R. A massive data framework for M-estimators with cubic-rate. *Journal of the American Statistical Association* 2018; 113(524): 1698–1709.

4. Jordan MI, Lee JD, Yang Y. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association* 2019; 114(526): 668–681.

5. Volgushev S, Chao SK, Cheng G. Distributed inference for quantile regression processes. *The Annals of Statistics* 2019; 47(3): 1634–1662.

6. Ma P, Mahoney M, Yu B. A statistical perspective on algorithmic leveraging. *Journal of Machine Learning Research* 2015; 16: 861–911.

7. Wang H, Zhu R, Ma P. Optimal subsampling for large sample Logistic regression. *Journal of the American Statistical Association* 2018; 113(522): 829–844.

8. Wang H. More efficient estimation for logistic regression with optimal subsample. *Journal of Machine Learning Research,* 2019; 20: 1–59.

9. Wang H, Yang M, Stufken J. Information-based optimal subdata selection for big data linear regression. *Journal of the American Statistical Association* 2019; 114(525): 393–405.

10. Wang H, Ma Y. Optimal subsampling for quantile regression in big data. *Biometrika* 2020: arXiv:2001.10168v1.

11. Zhang H, Wang H. Distributed subdata selection for big data via sampling-based approach. *Computational Statistics & Data Analysis* 2021; 153: 107072. doi: 10.1016/j.csda.2020.107072

12. Kawaguchi ES, Suchard MA, Liu Z, Li G. Scalable sparse Cox's regression for large-scale survival data via broken adaptive ridge.. 2018: arXiv:1712.00561v2.

13. Wang Y, Hong C, Palmer N, et al. A fast divide-and-conquer sparse Cox regression. *Biostatistics* 2019. doi: 10.1093/biostatistics/kxz036

14. Xue Y, Wang H, Yan J, Schifano ED. An online updating approach for testing the proportional hazards assumption with streams of survival data. *Biometrics* 2019; 76(1): 171–182.

15. Aalen OO. A linear regression model for the analysis of life times. *Statistics in Medicine* 1989; 8(8): 907–925.

16. Lin DY, Ying Z. Semiparametric analysis of the additive risk model. *Biometrika* 1994; 81(1): 61–71.

17. Kiefer J. Optimum experimental designs. *Journal of the Royal Statistical Society, Series B* 1959; 21: 272–319.

18. Mukhtar F, Boffetta P, Dabo B, et al. Disparities by race, age, and sex in the improvement of survival for lymphoma: Findings from a population-based study. *PLOS ONE* 2018; 13(7): e0199745.

19. Cox DR. Regression models and life-tables (with discussions). *Journal of the Royal Statistical Society, Series B* 1972; 34: 187–220.

20. Huang J, Ma S, Xie H. Regularized estimation in the accelerated failure time model with high-dimensional covariates. *Biometrics* 2006; 62(3): 813–820.

□

Supporting Information for "*Sampling-Based Estimation for Massive Survival Data with Additive Hazards Model*"

Lulu Zuo, Haixiang Zhang, HaiYing Wang and Lei Liu

The Supporting Information contains technical proofs for the theoretical results, as well as Tables S.1 − S.6 for the main document of this article. Moreover, an additional simulation study is presented. In order to characterize the asymptotic properties of the proposed subsample estimator, we need the following regularity assumptions:

(**A.1**) $\int_0^\tau \lambda_0(t)dt < \infty$.

(**A.2**) As $n \to \infty$, the $\dot{\Psi}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \int_0^\tau Y_i(t)\{\mathbf{X}_i - \bar{\mathbf{X}}(t)\}^{\otimes 2}dt$ converges to a positive definite matrix in probability.

(**A.3**) The parameter space $\Theta \subset \mathbb{R}^p$ is a compact convex set, and $\hat{\boldsymbol{\theta}}_{ZE}$ is in the interior of Θ.

(**A.4**) $\frac{1}{n}\sum_{i=1}^n \|\mathbf{X}_i\| = O_P(1)$ and $\frac{1}{n^2}\sum_{i=1}^n \frac{\|\mathbf{X}_i\|^2}{\pi_i} = O_p(1)$, where $\|\cdot\|$ denotes the Euclidean norm of a vector.

(**A.5**) $\sup_{\theta\in\Theta} \frac{1}{n^2}\sum_{i=1}^n \frac{\|\int_0^\tau\{\mathbf{X}_i - \bar{\mathbf{X}}(t)\}\{dN_i(t) - Y_i(t)\boldsymbol{\theta}'\mathbf{X}_i dt\}\|^2}{\pi_i} = O_p(1)$.

(**A.6**) $\frac{1}{n^2}\sum_{i=1}^n \frac{\{\int_0^\tau \|Y_i(t)\{\mathbf{X}_i - \bar{\mathbf{X}}(t)\}\|^2 dt\}^2}{\pi_i} = O_p(1)$.

(**A.7**) $\sup_{\theta\in\Theta} \frac{1}{n^3}\sum_{i=1}^n \frac{\|\int_0^\tau\{\mathbf{X}_i - \bar{\mathbf{X}}(t)\}\{dN_i(t) - Y_i(t)\boldsymbol{\theta}'\mathbf{X}_i dt\|^3}{\pi_i^2} = O_p(1)$.

(**A.8**) For $\boldsymbol{\theta} \in \Theta$ and $M_i(t) = N_i(t) - \int_0^t Y_i(s)\{d\Lambda_0(s) + \boldsymbol{\theta}'\mathbf{X}_i ds\}$, the $\frac{1}{n^2}\sum_{i=1}^n \frac{\{\int_0^\tau\{\mathbf{X}_i - \bar{\mathbf{X}}(t)\}dM_i(t)\}^{\otimes 2}}{\pi_i}$ and $\frac{1}{n^2}\sum_{i=1}^n \frac{\int_0^\tau\{\mathbf{X}_i - \bar{\mathbf{X}}(t)\}^{\otimes 2}dN_i(t)}{\pi_i}$ converge to a positive definite matrix in probability as $n \to \infty$, respectively.

Assumptions (A.1) and (A.2) are commonly used for the AH model, e.g. Lin and Ying (1994). Assumption (A.3) is a standard condition in the proofs. Assumptions (A.4), (A.5), (A.6) and (A.7) are some conditions on the subsampling probabilities and the AH model. For uniform subsampling with $\{\pi_i = 1/n\}_{i=1}^n$, sufficient conditions for these assumptions are $E\|\mathbf{X}\|^2 < \infty$, $E[\|\int_0^\tau\{\mathbf{X} - \bar{\mathbf{X}}(t)\}\{dN(t) - Y(t)\boldsymbol{\theta}'\mathbf{X}dt\}\|^2] < \infty$, $E\{\int_0^\tau Y(t)\|\mathbf{X} - \bar{\mathbf{X}}(t)\|^2 dt\}^2 < \infty$, and $E[\|\int_0^\tau\{\mathbf{X} - \bar{\mathbf{X}}(t)\}\{dN(t) - Y(t)\boldsymbol{\theta}'\mathbf{X}dt\}\|^3] < \infty$, respectively. Assumption (A.8) is used to establish the asymptotic properties of the subsample-based estimator. For uniform subsampling with $\{\pi_i = 1/n\}_{i=1}^n$, this assumption is reduced to the regularity condition in Lin and Ying (1994).

S.1 Proofs

We give the proof details for (6), (7) and (11) in the main text of our paper. First, we need the following lemmas.

Lemma 1 *If Assumptions (A.1), (A.3) and (A.4) hold, then as $n \to \infty$ and $r \to \infty$, conditionally on $\mathcal{F}_n$, for $\boldsymbol{\theta} \in \Theta$ we have*

$$\mathbf{U}^*(\boldsymbol{\theta}) = \Psi^*(\boldsymbol{\theta}) + o_{P|\mathcal{F}_n}(1), \tag{S.1}$$

and

$$\Psi^*(\boldsymbol{\theta}) = \Phi^*(\boldsymbol{\theta}) + o_{P|\mathcal{F}_n}(1), \tag{S.2}$$

where $\Psi^*(\boldsymbol{\theta}) = \frac{1}{nr}\sum_{i=1}^r \frac{1}{\pi_i^*}\int_0^\tau \{\mathbf{X}_i^* - \bar{\mathbf{X}}(t)\}\{dN_i^*(t) - Y_i^*(t)\boldsymbol{\theta}'\mathbf{X}_i^* dt\}$, $\Phi^*(\boldsymbol{\theta}) = \frac{1}{nr}\sum_{i=1}^r \frac{1}{\pi_i^*}\int_0^\tau \{\mathbf{X}_i^* - \bar{\mathbf{X}}(t)\}dM_i^*(t)$, and $M_i^*(t) = N_i^*(t) - \int_0^t Y_i^*(s)\{d\Lambda_0(s) + \boldsymbol{\theta}'\mathbf{X}_i^* ds\}$.

Proof. Note that

$$\begin{aligned}
\mathbf{U}^*(\boldsymbol{\theta}) &= \frac{1}{nr}\sum_{i=1}^r \frac{1}{\pi_i^*}\int_0^\tau \{\mathbf{X}_i^* - \bar{\mathbf{X}}^*(t)\}\{dN_i^*(t) - Y_i^*(t)\boldsymbol{\theta}'\mathbf{X}_i^* dt\} \\
&= \frac{1}{nr}\sum_{i=1}^r \frac{1}{\pi_i^*}\int_0^\tau \{\mathbf{X}_i^* - \bar{\mathbf{X}}(t) + \bar{\mathbf{X}}(t) - \bar{\mathbf{X}}^*(t)\}\{dN_i^*(t) - Y_i^*(t)\boldsymbol{\theta}'\mathbf{X}_i^* dt\} \\
&= \Psi^*(\boldsymbol{\theta}) + \frac{1}{nr}\sum_{i=1}^r \frac{1}{\pi_i^*}\int_0^\tau \{\bar{\mathbf{X}}(t) - \bar{\mathbf{X}}^*(t)\}\{dN_i^*(t) - Y_i^*(t)\boldsymbol{\theta}'\mathbf{X}_i^* dt\} \\
&= \Psi^*(\boldsymbol{\theta}) + \int_0^\tau \{\bar{\mathbf{X}}(t) - \bar{\mathbf{X}}^*(t)\}\frac{1}{nr}\sum_{i=1}^r \frac{1}{\pi_i^*}\{dN_i^*(t) - Y_i^*(t)\boldsymbol{\theta}'\mathbf{X}_i^* dt\}.
\end{aligned}$$

Given $\mathcal{F}_n$ and $t \in [0, \tau]$, direct calculation yields that

$$E\Big[\frac{1}{nr}\sum_{i=1}^r \frac{\{N_i^*(t) - Y_i^*(t)\boldsymbol{\theta}'\mathbf{X}_i^*\}}{\pi_i^*}\Big|\mathcal{F}_n\Big] = \frac{1}{n}\sum_{i=1}^n \{N_i(t) - Y_i(t)\boldsymbol{\theta}'\mathbf{X}_i\}, \tag{S.3}$$

and

$$\begin{aligned}
&E\Big[\frac{1}{nr}\sum_{i=1}^{r}\frac{\{N_i^*(t)-Y_i^*(t)\boldsymbol{\theta}'\mathbf{X}_i^*\}}{\pi_i^*}-\frac{1}{n}\sum_{i=1}^{n}\{N_i(t)-Y_i(t)\boldsymbol{\theta}'\mathbf{X}_i\}\Big|\mathcal{F}_n\Big]^2\\
&=\frac{1}{r}\Big[\frac{1}{n^2}\sum_{i=1}^{n}\frac{\{N_i(t)-Y_i(t)\boldsymbol{\theta}'\mathbf{X}_i\}^2}{\pi_i}-\Big(\frac{1}{n}\sum_{i=1}^{n}\{N_i(t)-Y_i(t)\boldsymbol{\theta}'\mathbf{X}_i\}\Big)^2\Big]\\
&\le\frac{1}{r}\Big[\frac{1}{n^2}\sum_{i=1}^{n}\frac{\{N_i(t)-Y_i(t)\boldsymbol{\theta}'\mathbf{X}_i\}^2}{\pi_i}\Big]\\
&\le\frac{1}{r}\Big[\frac{1}{n^2}\sum_{i=1}^{n}\frac{1+(\|\boldsymbol{\theta}\|\|\mathbf{X}_i\|)^2}{\pi_i}\Big]\\
&=O_{P|\mathcal{F}_n}(r^{-1}), \qquad \text{(S.4)}
\end{aligned}$$

where (S.4) holds by Assumptions (A.3) and (A.4). Using Markov's inequality, we have

$$\frac{1}{nr}\sum_{i=1}^{r}\frac{\{N_i^*(t)-Y_i^*(t)\boldsymbol{\theta}'\mathbf{X}_i^*\}}{\pi_i^*}=\frac{1}{n}\sum_{i=1}^{n}\{N_i(t)-Y_i(t)\boldsymbol{\theta}'\mathbf{X}_i\}+O_{P|\mathcal{F}_n}(r^{-1/2}). \qquad \text{(S.5)}$$

Similarly,

$$E\Big[\frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}Y_i^*(t)\mathbf{X}_i^*\Big|\mathcal{F}_n\Big]=\frac{1}{n}\sum_{i=1}^{n}Y_i(t)\mathbf{X}_i. \qquad \text{(S.6)}$$

From Assumption (A.4), we can derive that

$$E\Big[\frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}Y_i^*(t)\mathbf{X}_i^*-\frac{1}{n}\sum_{i=1}^{n}Y_i(t)\mathbf{X}_i\Big|\mathcal{F}_n\Big]^2\le\frac{1}{rn^2}\sum_{i=1}^{r}\frac{\|\mathbf{X}_i\|^2}{\pi_i}=O_{P|\mathcal{F}_n}(r^{-1}). \qquad \text{(S.7)}$$

Combining (S.6), (S.7) and Markov's inequality, we have

$$\frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}Y_i^*(t)\mathbf{X}_i^*=\frac{1}{n}\sum_{i=1}^{n}Y_i(t)\mathbf{X}_i+O_{P|\mathcal{F}_n}(r^{-1/2}). \qquad \text{(S.8)}$$

Similar to (S.8), we can deduce that

$$\frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}Y_i^*(t)=\frac{1}{n}\sum_{i=1}^{n}Y_i(t)+O_{P|\mathcal{F}_n}(r^{-1/2}). \qquad \text{(S.9)}$$

Note that

$$\bar{\mathbf{X}}^*(t)-\bar{\mathbf{X}}(t)=\frac{\frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}Y_i^*(t)\mathbf{X}_i^*\frac{1}{n}\sum_{i=1}^{n}Y_i(t)-\frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}Y_i^*(t)\frac{1}{n}\sum_{i=1}^{n}Y_i(t)\mathbf{X}_i}{\frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}Y_i^*(t)\frac{1}{n}\sum_{i=1}^{n}Y_i(t)}.$$

By Slutsky's theorem, Assumptions (A.4), (S.8) and (S.9), we know that $\bar{\mathbf{X}}(t)-\bar{\mathbf{X}}^*(t)=O_{P|\mathcal{F}_n}(r^{-1/2})$. Denote $f(t)=\frac{1}{nr}\sum_{i=1}^r\frac{1}{\pi_i^*}\{N_i^*(t)-Y_i^*(t)\boldsymbol{\theta}'\mathbf{X}_i^*\}$. From Royden and Fitzpatrick (2010), we know that $f=f^+(t)-f^-(t)$, where $f^+(t)$ and $f^-(t)$ are positive and monotone functions of t. In view of (S.5) and Lemma 1 of Lin et al. (2000), we can ensure that as $n\to\infty$ and $r\to\infty$,

$$\begin{aligned}&\int_0^\tau\{\bar{\mathbf{X}}(t)-\bar{\mathbf{X}}^*(t)\}\frac{1}{nr}\sum_{i=1}^r\frac{1}{\pi_i^*}\{dN_i^*(t)-Y_i^*(t)\boldsymbol{\theta}'\mathbf{X}_i^*dt\}\\&=\int_0^\tau\{\bar{\mathbf{X}}(t)-\bar{\mathbf{X}}^*(t)\}df^+(t)-\int_0^\tau\{\bar{\mathbf{X}}(t)-\bar{\mathbf{X}}^*(t)\}df^-(t)\\&=o_{P|\mathcal{F}_n}(1).\end{aligned}\tag{S.10}$$

Hence, (S.1) holds. Next we prove (S.2). Note that

$$\begin{aligned}\Psi^*(\boldsymbol{\theta})&=\Phi^*(\boldsymbol{\theta})+\frac{1}{nr}\sum_{i=1}^r\frac{1}{\pi_i^*}\int_0^\tau\{\mathbf{X}_i^*-\bar{\mathbf{X}}(t)\}Y_i^*(t)\lambda_0(t)dt\\&=\Phi^*(\boldsymbol{\theta})+\int_0^\tau\frac{1}{nr}\sum_{i=1}^r\frac{1}{\pi_i^*}\{\mathbf{X}_i^*-\bar{\mathbf{X}}(t)\}Y_i^*(t)\lambda_0(t)dt\\&=\Phi^*(\boldsymbol{\theta})+\int_0^\tau\left[\frac{1}{nr}\sum_{i=1}^r\frac{1}{\pi_i^*}\mathbf{X}_i^*Y_i^*(t)-\bar{\mathbf{X}}(t)\left\{\frac{1}{nr}\sum_{i=1}^r\frac{1}{\pi_i^*}Y_i^*(t)\right\}\right]\lambda_0(t)dt.\end{aligned}$$

Given $\mathcal{F}_n$ and $t\in[0,\tau]$, the (S.8) and (S.9) lead to

$$\frac{1}{nr}\sum_{i=1}^r\frac{1}{\pi_i^*}\mathbf{X}_i^*Y_i^*(t)-\bar{\mathbf{X}}(t)\left\{\frac{1}{nr}\sum_{i=1}^r\frac{1}{\pi_i^*}Y_i^*(t)\right\}=O_{P|\mathcal{F}_n}(r^{-1/2}).\tag{S.11}$$

By mean value theorem for integrals, together with Assumption (A.1) and (S.11), as $r\to\infty$ we can get

$$\int_0^\tau\Big[\frac{1}{nr}\sum_{i=1}^r\frac{1}{\pi_i^*}\mathbf{X}_i^*Y_i^*(t)-\bar{\mathbf{X}}(t)\frac{1}{nr}\sum_{i=1}^r\frac{1}{\pi_i^*}Y_i^*(t)\Big]\lambda_0(t)dt=o_{P|\mathcal{F}_n}(1).\tag{S.12}$$

□

Lemma 2 *If Assumptions (A.1)– (A.7) hold, then as $n\to\infty$ and $r\to\infty$, conditionally on $\mathcal{F}_n$, we have*

$$\Psi^*(\hat{\boldsymbol{\theta}}_{ZE})=O_{P|\mathcal{F}_n}(r^{-1/2}),\tag{S.13}$$

and

$$\dot{\Psi}^*(\hat{\boldsymbol{\theta}}_{ZE})^{-1} = O_{P|\mathcal{F}_n}(1), \tag{S.14}$$

where $\dot{\Psi}^*(\hat{\boldsymbol{\theta}}_{ZE}) = \frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}\int_0^\tau Y_i^*(t)\{\mathbf{X}_i^* - \bar{\mathbf{X}}(t)\}\mathbf{X}_i^{*\prime}dt.$

Proof. For $\boldsymbol{\theta} \in \Theta$, we can derive that

$$E\{\Psi^*(\boldsymbol{\theta})|\mathcal{F}_n\} = \Psi(\boldsymbol{\theta}). \tag{S.15}$$

For the jth component of $\Psi^*(\boldsymbol{\theta})$, denoted as $\Psi_j^*(\boldsymbol{\theta}) = \frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}\psi_{ij}^*(\boldsymbol{\theta})$, where $\psi_{ij}^*(\boldsymbol{\theta}) = \int_0^\tau\{\mathbf{X}_{ij}^* - \bar{\mathbf{X}}_j(t)\}\{dN_i^*(t) - Y_i^*(t)\boldsymbol{\theta}'\mathbf{X}_i^*dt\}$, we have

$$\begin{aligned}
E\big\{\Psi_j^*(\boldsymbol{\theta}) - \Psi_j(\boldsymbol{\theta})|\mathcal{F}_n\big\}^2 &= E\Big\{\frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}\psi_{ij}^*(\boldsymbol{\theta}) - \frac{1}{n}\sum_{l=1}^{n}\psi_{lj}(\boldsymbol{\theta})\Big|\mathcal{F}_n\Big\}^2 \\
&= \frac{1}{n^2r^2}E\Big[\sum_{i=1}^{r}\Big\{\frac{\psi_{ij}^*(\boldsymbol{\theta})}{\pi_i^*} - \sum_{l=1}^{n}\psi_{lj}(\boldsymbol{\theta})\Big\}\Big|\mathcal{F}_n\Big]^2 \\
&= \frac{1}{n^2r^2}E\Big[\sum_{i=1}^{r}\Big\{\frac{\psi_{ij}^*(\boldsymbol{\theta})}{\pi_i^*} - \sum_{l=1}^{n}\psi_{lj}(\boldsymbol{\theta})\Big\}^2 \\
&+ \sum_{i\neq k}\Big\{\frac{\psi_{ij}^*(\boldsymbol{\theta})}{\pi_i^*} - \sum_{l=1}^{n}\psi_{lj}(\boldsymbol{\theta})\Big\}\Big\{\frac{\psi_{kj}^*(\boldsymbol{\theta})}{\pi_k^*} - \sum_{l=1}^{n}\psi_{lj}(\boldsymbol{\theta})\Big\}\Big|\mathcal{F}_n\Big] \\
&= \frac{1}{n^2r^2}E\Big[\sum_{i=1}^{r}\Big\{\frac{\psi_{ij}^*(\boldsymbol{\theta})}{\pi_i^*} - \sum_{l=1}^{n}\psi_{lj}(\boldsymbol{\theta})\Big\}^2\Big|\mathcal{F}_n\Big] \\
&= \frac{1}{n^2r}\cdot\sum_{i=1}^{n}\Big\{\frac{\psi_{ij}(\boldsymbol{\theta})}{\pi_i} - \sum_{l=1}^{n}\psi_{lj}(\boldsymbol{\theta})\Big\}^2\cdot\pi_i \\
&= \frac{1}{r}\Big[\frac{1}{n^2}\sum_{i=1}^{n}\frac{\psi_{ij}(\boldsymbol{\theta})^2}{\pi_i} - \Big\{\frac{1}{n}\sum_{i=1}^{n}\psi_{ij}(\boldsymbol{\theta})\Big\}^2\Big] \\
&\leq \frac{1}{r}\Big[\frac{1}{n^2}\sum_{i=1}^{n}\frac{\|\int_0^\tau\{\mathbf{X}_i - \bar{\mathbf{X}}(t)\}\{dN_i(t) - Y_i(t)\boldsymbol{\theta}'\mathbf{X}_idt\}\|^2}{\pi_i} \\
&- \Big\{\frac{1}{n}\sum_{i=1}^{n}\Big\|\int_0^\tau\{\mathbf{X}_i - \bar{\mathbf{X}}(t)\}\{dN_i(t) - Y_i(t)\boldsymbol{\theta}'\mathbf{X}_idt\}\Big\|\Big\}^2\Big] \\
&\leq \frac{1}{rn^2}\sum_{i=1}^{n}\frac{\|\int_0^\tau\{\mathbf{X}_i - \bar{\mathbf{X}}(t)\}\{dN_i(t) - Y_i(t)\boldsymbol{\theta}'\mathbf{X}_idt\}\|^2}{\pi_i}.
\end{aligned}$$

By Assumption (A.5), we have

$$E\big\{\Psi_j^*(\boldsymbol{\theta}) - \Psi_j(\boldsymbol{\theta})|\mathcal{F}_n\big\}^2 = O_{P|\mathcal{F}_n}(r^{-1}).$$

From Markov's inequality, together with (S.15), we can get

$$\Psi^*(\boldsymbol{\theta}) - \Psi(\boldsymbol{\theta}) = O_{P|\mathcal{F}_n}(r^{-1/2}). \tag{S.16}$$

By Assumption (A.3), we have $\Psi^*(\hat{\boldsymbol{\theta}}_{ZE}) - \Psi(\hat{\boldsymbol{\theta}}_{ZE}) = O_{P|\mathcal{F}_n}(r^{-1/2})$. Because $\Psi(\hat{\boldsymbol{\theta}}_{ZE}) = 0$, it follows that (S.13) holds.

To prove (S.14), some direct calculations yield that

$$E\{\dot{\Psi}^*(\boldsymbol{\theta})|\mathcal{F}_n\} = \dot{\Psi}(\boldsymbol{\theta}).$$

For any component $\dot{\Psi}^*_{j_1j_2}(\boldsymbol{\theta})$ of $\dot{\Psi}^*(\boldsymbol{\theta})$, where $1 \le j_1, j_2 \le p$, we can derive that

$$\begin{aligned}
&E\big\{\dot{\Psi}^*_{j_1j_2}(\boldsymbol{\theta}) - \dot{\Psi}_{j_1j_2}(\boldsymbol{\theta})|\mathcal{F}_n\big\}^2 \\
&= \frac{1}{r}\left[\frac{1}{n^2}\sum_{i=1}^n \frac{\{\int_0^\tau Y_i(t)\{\mathbf{X}_i - \bar{\mathbf{X}}(t)\}^{\otimes 2}dt\}^2_{j_1j_2}}{\pi_i} - \left\{\frac{1}{n}\sum_{i=1}^n \Big(\int_0^\tau Y_i(t)\{\mathbf{X}_i - \bar{\mathbf{X}}(t)\}^{\otimes 2}dt\Big)_{j_1j_2}\right\}^2\right] \\
&\le \frac{1}{rn^2}\sum_{i=1}^n \frac{\{\int_0^\tau Y_i(t)\{\mathbf{X}_i - \bar{\mathbf{X}}(t)\}^{\otimes 2}dt\}^2_{j_1j_2}}{\pi_i} \\
&\le \frac{1}{rn^2}\sum_{i=1}^n \frac{\{\int_0^\tau \|Y_i(t)\{\mathbf{X}_i - \bar{\mathbf{X}}(t)\}\|^2dt\}^2}{\pi_i}.
\end{aligned}$$

By Assumption (A.6), we have

$$E\big\{\dot{\Psi}^*_{j_1j_2}(\boldsymbol{\theta}) - \dot{\Psi}_{j_1j_2}(\boldsymbol{\theta})|\mathcal{F}_n\big\}^2 = O_{P|\mathcal{F}_n}(r^{-1/2}). \tag{S.17}$$

From Markov's inequality, we can derive the following equation

$$\dot{\Psi}^*(\boldsymbol{\theta}) - \dot{\Psi}(\boldsymbol{\theta}) = O_{P|\mathcal{F}_n}(r^{-1/2}).$$

Based on Assumptions (A.2) and (A.3), we know that (S.14) holds. □

Proof of (6). As $r \to \infty$, Lemma 1 and (S.16) lead to $\mathbf{U}^*(\boldsymbol{\theta}) - \Psi(\boldsymbol{\theta}) \to 0$ in probability conditional on $\mathcal{F}_n$. Note that the parameter space is compact, and $\hat{\boldsymbol{\theta}}_{ZE}$ is the unique solution of $\Psi(\boldsymbol{\theta}) = 0$ (Lin and Ying , 1994). Thus, from Theorem 5.9 and its remark of van der Vaart (1998), conditionally on $\mathcal{F}_n$, as $n \to \infty$ and $r \to \infty$, we know that

$$\|\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{ZE}\| = o_{P|\mathcal{F}_n}(1).$$

Using the Taylor's theorem (Ferguson, 1996, Chapter 4) and (S.1) of Lemma 1, we have

$$0 = \mathbf{U}_j^*(\tilde{\boldsymbol{\theta}}) = \Psi_j^*(\hat{\boldsymbol{\theta}}_{ZE}) + \dot{\Psi}_j^*(\hat{\boldsymbol{\theta}}_{ZE})(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{ZE}) + o_{P|\mathcal{F}_n}(1). \tag{S.18}$$

Then,

$$\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{ZE} = -\dot{\Psi}^*(\hat{\boldsymbol{\theta}}_{ZE})^{-1}\Psi^*(\hat{\boldsymbol{\theta}}_{ZE}) + o_{P|\mathcal{F}_n}(1). \tag{S.19}$$

From Lemma 2, it is obvious that $\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{ZE} = O_{P|\mathcal{F}_n}(r^{-1/2})$. $\square$

Proof of (7). Note that

$$\Phi^*(\hat{\boldsymbol{\theta}}_{ZE}) = \frac{1}{r}\sum_{i=1}^{r} \frac{\int_0^\tau \{\mathbf{X}_i^* - \bar{\mathbf{X}}(t)\} dM_i^*(t)}{n\pi_i^*} = \frac{1}{r}\sum_{i=1}^{r} W_i^*, \tag{S.20}$$

where $W_i^* = \frac{\int_0^\tau \{\mathbf{X}_i^* - \bar{\mathbf{X}}(t)\} dM_i^*(t)}{n\pi_i^*}$, and $M_i^*(t)$ is defined in Lemma 1. Given $\mathcal{F}_n$, we know that $W_1^*, \cdots, W_r^*$ are independently and identically distributed random variables with

$$\begin{aligned} E(W_i^*|\mathcal{F}_n) &= E\left[\frac{\int_0^\tau \{\mathbf{X}_i^* - \bar{\mathbf{X}}(t)\} dM_i^*(t)}{n\pi_i^*}\bigg|\mathcal{F}_n\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \int_0^\tau \{\mathbf{X}_i - \bar{\mathbf{X}}(t)\}\{dN_i(t) - Y_i(t)\hat{\boldsymbol{\theta}}_{ZE}'\mathbf{X}_i dt\} \\ &= \Psi(\hat{\boldsymbol{\theta}}_{ZE}) = 0, \end{aligned}$$

and

$$\begin{aligned} Var\{W_i^*|\mathcal{F}_n\} &= E\left[\frac{1}{n^2\pi_i^{*2}}\Big\{\int_0^\tau \{\mathbf{X}_i^* - \bar{\mathbf{X}}(t)\} dM_i^*(t)\Big\}^{\otimes 2}\bigg|\mathcal{F}_n\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \frac{\{\int_0^\tau \{\mathbf{X}_i - \bar{\mathbf{X}}(t)\} dM_i(t)\}^{\otimes 2}}{\pi_i}, \end{aligned} \tag{S.21}$$

where $M_i(t) = N_i(t) - \int_0^t Y_i(s)\{d\Lambda_0(s) + \boldsymbol{\theta}'\mathbf{X}_i ds\}$. Note that $M_i(\cdot)$ is a local squared integrable martingale, and the variation process $\langle M_i, M_i\rangle(t)$ satisfies (Fleming and Harrington, 1991)

$$\langle M_i, M_i\rangle(t) = \Lambda_i(t) = \int_0^t Y_i(s)\{d\Lambda_0(s) + \boldsymbol{\theta}'\mathbf{X}_i ds\}, \quad \text{and} \quad \langle M_i, M_j\rangle(t) = 0, i \neq j. \tag{S.22}$$

From Theorem 2.4.1 and Lemma 2.4.1 of Fleming and Harrington (1991), for any $t \in [0, \tau]$,

$$\begin{aligned}
&E\Big[\frac{1}{n^2}\sum_{i=1}^n \frac{\{\int_0^t \{\mathbf{X}_i - \bar{\mathbf{X}}(s)\}dM_i(s)\}^{\otimes 2}}{\pi_i}\Big] \\
&= E\Big[\frac{1}{n^2}\sum_{i=1}^n \frac{\int_0^t \{\mathbf{X}_i - \bar{\mathbf{X}}(s)\}^{\otimes 2} d\langle M_i, M_i\rangle(s)}{\pi_i}\Big] \\
&= E\Big[\frac{1}{n^2}\sum_{i=1}^n \frac{\int_0^t \{\mathbf{X}_i - \bar{\mathbf{X}}(s)\}^{\otimes 2} d\Lambda_i(s)}{\pi_i}\Big] \\
&= E\Big[\frac{1}{n^2}\sum_{i=1}^n \frac{\int_0^t \{\mathbf{X}_i - \bar{\mathbf{X}}(s)\}^{\otimes 2} dN_i(s)}{\pi_i}\Big].
\end{aligned} \tag{S.23}$$

By Assumption (A.8) and (S.23), as $n \to \infty$ we have

$$\frac{1}{n^2}\sum_{i=1}^n \frac{\{\int_0^\tau \{\mathbf{X}_i - \bar{\mathbf{X}}(t)\}dM_i(t)\}^{\otimes 2}}{\pi_i} = \frac{1}{n^2}\sum_{i=1}^n \frac{\int_0^\tau \{\mathbf{X}_i - \bar{\mathbf{X}}(t)\}^{\otimes 2} dN_i(t)}{\pi_i} + o_P(1). \tag{S.24}$$

Thus,

$$Var\{W_i^* | \mathcal{F}_n\} = \frac{1}{n^2}\sum_{i=1}^n \frac{\int_0^\tau \{\mathbf{X}_i - \bar{\mathbf{X}}(t)\}^{\otimes 2} dN_i(t)}{\pi_i} + o_{P|\mathcal{F}_n}(1). \tag{S.25}$$

By Assumption (A.8), it is known that $Var(W_i^* | \mathcal{F}_n) = O_P(1)$ as $n \to \infty$. Meanwhile, for every $\varepsilon > 0$,

$$\begin{aligned}
&E\Big\{\sum_{i=1}^r (r^{-1/2}W_i^*)^2 I(|W_i^* r^{-1/2}| > \varepsilon)\Big|\mathcal{F}_n\Big\} \\
&\leq \frac{1}{r}\sum_{i=1}^r E\Big(\|W_i^*\|^2 \cdot \frac{\|W_i^*\|}{r^{1/2}\varepsilon}\Big|\mathcal{F}_n\Big) \\
&= \frac{1}{r^{1/2}\varepsilon} E(\|W_i^*\|^3 | \mathcal{F}_n) \\
&= \frac{1}{r^{1/2}\varepsilon} \cdot \frac{1}{n^3}\sum_{i=1}^n \frac{\|\int_0^\tau \{\mathbf{X}_i - \bar{\mathbf{X}}(t)\}\{dN_i(t) - Y_i(t)\boldsymbol{\theta}'\mathbf{X}_i dt\}\|^3}{\pi_i^2}.
\end{aligned}$$

By Assumption (A.7), as $r \to \infty$ we have

$$E\Big\{\sum_{i=1}^{r}(r^{-1/2}W_i^*)^2 I(|W_i^* r^{-1/2}| > \varepsilon)\Big|\mathcal{F}_n\Big\} \leq \frac{1}{r^{1/2}\varepsilon}O_P(1) = o_p(1). \tag{S.26}$$

From (S.20) and (S.25), together with the Lindeberg-Feller central limit theorem (Proposition 2.27 of van der Vaart, 1998) and the Slutsky's theorem, conditionally on $\mathcal{F}_n$, it can be proved that as $n \to \infty$ and $r \to \infty$, $\mathbf{\Gamma}^{-1/2}\Phi^*(\hat{\boldsymbol{\theta}}_{ZE}) \xrightarrow{d} N(0, \mathbf{I})$. By Lemma 1 and the Slutsky's theorem, we have

$$\mathbf{\Gamma}^{-1/2}\Psi^*(\hat{\boldsymbol{\theta}}_{ZE}) \xrightarrow{d} N(0, \mathbf{I}). \tag{S.27}$$

Based on Lemma 2, (S.19) and Theorem 1, we can get that

$$\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{ZE} = -\{\dot{\Psi}^*(\hat{\boldsymbol{\theta}}_{ZE})\}^{-1}\Psi^*(\hat{\boldsymbol{\theta}}_{ZE}) + o_{P|\mathcal{F}_n}(1). \tag{S.28}$$

Note that

$$\begin{aligned}
&-\dot{\Psi}^*(\hat{\boldsymbol{\theta}}_{ZE})^{-1}\Psi^*(\hat{\boldsymbol{\theta}}_{ZE}) \\
&= -\dot{\Psi}(\hat{\boldsymbol{\theta}}_{ZE})^{-1}\Psi^*(\hat{\boldsymbol{\theta}}_{ZE}) - \Big(\dot{\Psi}^*(\hat{\boldsymbol{\theta}}_{ZE})^{-1} - \dot{\Psi}(\hat{\boldsymbol{\theta}}_{ZE})^{-1}\Big)\Psi^*(\hat{\boldsymbol{\theta}}_{ZE}) \\
&= -\dot{\Psi}(\hat{\boldsymbol{\theta}}_{ZE})^{-1}\Psi^*(\hat{\boldsymbol{\theta}}_{ZE}) + \Big[\dot{\Psi}(\hat{\boldsymbol{\theta}}_{ZE})^{-1}\Big\{\dot{\Psi}^*(\hat{\boldsymbol{\theta}}_{ZE}) - \dot{\Psi}(\hat{\boldsymbol{\theta}}_{ZE})\Big\}\dot{\Psi}^*(\hat{\boldsymbol{\theta}}_{ZE})^{-1}\Big]\Psi^*(\hat{\boldsymbol{\theta}}_{ZE}) \\
&= -\dot{\Psi}(\hat{\boldsymbol{\theta}}_{ZE})^{-1}\Psi^*(\hat{\boldsymbol{\theta}}_{ZE}) + O_{P|\mathcal{F}_n}(1)O_{P|\mathcal{F}_n}(r^{-1/2})O_{P|\mathcal{F}_n}(1)O_{P|\mathcal{F}_n}(r^{-1/2}) \\
&= -\dot{\Psi}(\hat{\boldsymbol{\theta}}_{ZE})^{-1}\Psi^*(\hat{\boldsymbol{\theta}}_{ZE}) + O_{P|\mathcal{F}_n}(r^{-1}).
\end{aligned}$$

Hence,

$$\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{ZE} = -\dot{\Psi}(\hat{\boldsymbol{\theta}}_{ZE})^{-1}\Psi^*(\hat{\boldsymbol{\theta}}_{ZE}) + O_{P|\mathcal{F}_n}(r^{-1}). \tag{S.29}$$

By Assumptions (A.2) and (A.8), together with the fact that $\dot{\Psi}(\hat{\boldsymbol{\theta}}_{ZE}) = \mathcal{H}_X$, we get

$$\mathbf{\Sigma} = \mathcal{H}_X^{-1}\mathbf{\Gamma}\mathcal{H}_X^{-1} = \{\dot{\Psi}(\hat{\boldsymbol{\theta}}_{ZE})\}^{-1}\mathbf{\Gamma}\{\dot{\Psi}(\hat{\boldsymbol{\theta}}_{ZE})\}^{-1} = O_{P|\mathcal{F}_n}(r^{-1}). \tag{S.30}$$

Thus, (S.29) and (S.30) yield that

$$\mathbf{\Sigma}^{-1/2}(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{ZE}) = -\mathbf{\Sigma}^{-1/2}\mathcal{H}_X^{-1}\Psi^*(\hat{\boldsymbol{\theta}}_{ZE}) + O_{P|\mathcal{F}_n}(r^{-1/2})$$

$$= \quad -\mathbf{\Sigma}^{-1/2}\mathcal{H}_X^{-1}\mathbf{\Gamma}^{1/2}\mathbf{\Gamma}^{-1/2}\Psi^*(\hat{\boldsymbol{\theta}}_{ZE}) + O_{P|\mathcal{F}_n}(r^{-1/2}). \qquad \text{(S.31)}$$

Note that

$$\mathbf{\Sigma}^{-1/2}\mathcal{H}_X^{-1}\mathbf{\Gamma}^{1/2}(\mathbf{\Sigma}^{-1/2}\mathcal{H}_X^{-1}\mathbf{\Gamma}^{1/2})' = \mathbf{\Sigma}^{-1/2}\mathcal{H}_X^{-1}\mathbf{\Gamma}^{1/2}\mathbf{\Gamma}^{1/2}\mathcal{H}_X^{-1}\mathbf{\Sigma}^{-1/2} = \mathbf{I}. \qquad \text{(S.32)}$$

By (S.31), (S.32) and the Slutsky's theorem, as $n \to \infty$ and $r \to \infty$, we have

$$\mathbf{\Sigma}^{-1/2}(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{ZE}) \xrightarrow{d} N(0, \mathbf{I}).$$

□

Proof of (11). It can be deduced that

$$\begin{aligned}
tr(\mathbf{\Gamma}) &= tr\left(\frac{1}{rn^2}\sum_{i=1}^{n}\frac{\int_0^\tau\{\mathbf{X}_i - \bar{\mathbf{X}}(t)\}^{\otimes 2}dN_i(t)}{\pi_i}\right) \\
&= \frac{1}{rn^2}\sum_{i=1}^{n}\frac{tr(\int_0^\tau\{\mathbf{X}_i - \bar{\mathbf{X}}(t)\}^{\otimes 2}dN_i(t))}{\pi_i} \\
&= \frac{1}{rn^2}\Bigg[\sum_{i\in S_0}\frac{tr(\int_0^\tau\{\mathbf{X}_i - \bar{\mathbf{X}}(t)\}^{\otimes 2}dN_i(t))}{\pi_i} + \sum_{i\in S_1}\frac{tr(\int_0^\tau\{\mathbf{X}_i - \bar{\mathbf{X}}(t)\}^{\otimes 2}dN_i(t))}{\pi_i}\Bigg] \\
&= \frac{1}{rn^2(1-\delta)}\sum_{i\in S_1}\pi_i\sum_{i\in S_1}\frac{tr(\int_0^\tau\{\mathbf{X}_i - \bar{\mathbf{X}}(t)\}^{\otimes 2}dN_i(t))}{\pi_i} \qquad \text{(S.33)} \\
&\geq \frac{1}{rn^2(1-\delta)}\Bigg[\sum_{i\in S_1}tr^{1/2}\Big\{\int_0^\tau\{\mathbf{X}_i - \bar{\mathbf{X}}(t)\}^{\otimes 2}dN_i(t)\Big\}\Bigg]^2, \qquad \text{(S.34)}
\end{aligned}$$

where (S.33) holds by the fact that $dN_i(t) = 0$ for $i \in S_0$. Hence, (S.34) follows by the Cauchy-Schwarz inequality. The equality in (S.34) holds if and only if $\pi_i \propto tr^{1/2}\{\int_0^\tau\{\mathbf{X}_i - \bar{\mathbf{X}}(t)\}^{\otimes 2}dN_i(t)\}$ for $i \in S_1$. □

References

Ferguson, T. (1996). *A Course in Large Sample Theory.* New York: Chapman and Hall.

Fleming, T. and Harrington, D. (1991). *Counting Processes and Survival Analysis.* New York: John Wiley and Sons.

Lin, D. Y. and Ying, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika*, **81**, 61-71.

Lin, D. Y., Wei, L. J., Yang, I. and Ying, Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society Series B*, **62**, 711-730.

Royden, H. and Fitzpatrick, P. (2010). *Real Analysis* (Fourth Edition). Prentice Hall, Boston.

van der Vaart, A. (1998). *Asymptotic Statistics.* London: Cambridge University Press.

Table S.1 Simulation results on the subsample estimator with Case II[†].

		OSP				UNIF			
	r	bias	ESE	SSE	CP	bias	ESE	SSE	CP
$\theta_1 = -1$	100	0.0486	0.2631	0.2600	0.951	0.0576	0.2705	0.2687	0.947
	300	0.0164	0.1417	0.1354	0.965	0.0256	0.1457	0.1451	0.956
	500	0.0120	0.1087	0.1077	0.958	0.0108	0.1108	0.1071	0.959
$\theta_2 = -0.5$	100	0.0217	0.2002	0.1967	0.959	0.0364	0.2096	0.2088	0.963
	300	0.0049	0.1057	0.1043	0.962	0.0129	0.1090	0.1090	0.954
	500	0.0058	0.0805	0.0816	0.952	0.0073	0.0829	0.0838	0.952
$\theta_3 = 0$	100	0.0016	0.1765	0.1789	0.955	0.0020	0.1830	0.1896	0.955
	300	0.0022	0.0919	0.0969	0.940	0.0014	0.0937	0.0957	0.940
	500	0.0014	0.0693	0.0686	0.958	0.0045	0.0714	0.0719	0.947
$\theta_4 = 0.5$	100	0.0125	0.1999	0.2007	0.959	0.0271	0.2065	0.2021	0.968
	300	0.0119	0.1059	0.1040	0.954	0.0113	0.1088	0.1111	0.947
	500	0.0047	0.0807	0.0822	0.949	0.0049	0.0826	0.0813	0.958
$\theta_5 = 1$	100	0.0486	0.2597	0.2617	0.954	0.0622	0.2703	0.2662	0.953
	300	0.0142	0.1400	0.1317	0.967	0.0235	0.1450	0.1412	0.954
	500	0.0093	0.1076	0.1068	0.947	0.0090	0.1106	0.1084	0.958

[†] "OSP" denotes the proposed method with optimal subsampling probabilities; "UNIF" denotes the proposed method with uniform subsampling probabilities; "bias" denotes the sample mean of the estimates minus the full data estimator $\hat{\boldsymbol{\theta}}_{ZE}$; "ESE" denotes the estimated standard error of the estimates; "SSE" denotes the sampling standard error of the estimates; "CP" denotes the empirical 95% coverage probability towards $\hat{\boldsymbol{\theta}}_{ZE}$.

Table S.2 Simulation results on the subsample estimator with Case III[†].

		OSP				UNIF			
	r	bias	ESE	SSE	CP	bias	ESE	SSE	CP
$\theta_1 = -1$	100	0.0374	0.3551	0.3793	0.950	0.0345	0.3616	0.3852	0.950
	300	0.0098	0.1808	0.1794	0.955	0.0149	0.1868	0.1888	0.951
	500	0.0057	0.1360	0.1299	0.970	0.0038	0.1399	0.1422	0.950
$\theta_2 = -0.5$	100	0.0640	0.4066	0.4174	0.965	0.0525	0.4457	0.4358	0.963
	300	0.0215	0.2142	0.2253	0.943	0.0153	0.2379	0.2440	0.946
	500	0.0103	0.1632	0.1662	0.944	0.0044	0.1806	0.1769	0.958
$\theta_3 = 0$	100	0.0646	0.3425	0.3664	0.956	0.0553	0.3690	0.3915	0.951
	300	0.0190	0.1738	0.1761	0.944	0.0167	0.1892	0.1983	0.938
	500	0.0124	0.1311	0.1396	0.942	0.0090	0.1417	0.1453	0.942
$\theta_4 = 0.5$	100	0.0669	0.4137	0.4310	0.954	0.0631	0.4461	0.4480	0.954
	300	0.0215	0.2168	0.2166	0.951	0.0340	0.2415	0.2445	0.951
	500	0.0214	0.1662	0.1744	0.932	0.0185	0.1822	0.1781	0.950
$\theta_5 = 1$	100	0.0876	0.4724	0.4711	0.963	0.1278	0.5294	0.5287	0.955
	300	0.0279	0.2533	0.2648	0.953	0.0446	0.2833	0.2814	0.948
	500	0.0095	0.1936	0.1983	0.949	0.0164	0.2153	0.2091	0.961

[†] "OSP" denotes the proposed method with optimal subsampling probabilities; "UNIF" denotes the proposed method with uniform subsampling probabilities; "bias" denotes the sample mean of the estimates minus the full data estimator $\hat{\boldsymbol{\theta}}_{ZE}$; "ESE" denotes the estimated standard error of the estimates; "SSE" denotes the sampling standard error of the estimates; "CP" denotes the empirical 95% coverage probability towards $\hat{\boldsymbol{\theta}}_{ZE}$.

Table S.3 Simulation results on the subsample estimator with Case IV†.

		OSP				UNIF			
	r	bias	ESE	SSE	CP	bias	ESE	SSE	CP
$\theta_1 = -1$	100	0.0465	0.2632	0.2614	0.959	0.0796	0.2920	0.2967	0.950
	300	0.0233	0.1420	0.1415	0.953	0.0210	0.1558	0.1518	0.954
	500	0.0097	0.1081	0.1077	0.943	0.0041	0.1188	0.1173	0.956
$\theta_2 = -0.5$	100	0.0250	0.2193	0.2223	0.962	0.0287	0.2360	0.2363	0.957
	300	0.0157	0.1132	0.1103	0.959	0.0089	0.1229	0.1159	0.965
	500	0.0015	0.0860	0.0835	0.956	0.0033	0.0936	0.0901	0.962
$\theta_3 = 0$	100	0.0095	0.1936	0.1876	0.961	0.0019	0.2094	0.2066	0.964
	300	0.0038	0.0985	0.0955	0.956	0.0013	0.1066	0.1071	0.949
	500	0.0001	0.0740	0.0715	0.956	0.0003	0.0807	0.0825	0.949
$\theta_4 = 0.5$	100	0.0204	0.2183	0.2181	0.956	0.0328	0.2382	0.2367	0.965
	300	0.0147	0.1128	0.1116	0.956	0.0127	0.1238	0.1291	0.947
	500	0.0016	0.0849	0.0869	0.943	0.0050	0.0940	0.0891	0.963
$\theta_5 = 1$	100	0.0615	0.2632	0.2739	0.951	0.0742	0.2911	0.2905	0.961
	300	0.0146	0.1404	0.1405	0.953	0.0209	0.1556	0.1563	0.952
	500	0.0124	0.1074	0.1090	0.955	0.0093	0.1188	0.1173	0.957

† "OSP" denotes the proposed method with optimal subsampling probabilities; "UNIF" denotes the proposed method with uniform subsampling probabilities; "bias" denotes the sample mean of the estimates minus the full data estimator $\hat{\boldsymbol{\theta}}_{ZE}$; "ESE" denotes the estimated standard error of the estimates; "SSE" denotes the sampling standard error of the estimates; "CP" denotes the empirical 95% coverage probability towards $\hat{\boldsymbol{\theta}}_{ZE}$.

Table S.4

Simulation results on OSP-based $\breve{\theta}_1$ under varying censoring rates (Case II)†.

	δ	bias	ESE	SSE	CP
$r = 100$	16%	0.0480	0.2479	0.2440	0.957
	28%	0.0486	0.2631	0.2600	0.951
	38%	0.0518	0.3027	0.3047	0.952
$r = 300$	16%	0.0057	0.1337	0.1265	0.956
	28%	0.0164	0.1417	0.1354	0.965
	38%	0.0172	0.1636	0.1612	0.953
$r = 500$	16%	0.0065	0.1024	0.1013	0.957
	28%	0.0120	0.1087	0.1077	0.958
	38%	0.0065	0.1251	0.1233	0.958

† δ is the censoring rate; "bias", "ESE", "SSE" and "CP" are given in the footnotes of Table S.1.

Table S.5

Simulation results on OSP-based $\breve{\theta}_1$ under varying censoring rate (Case III)[†].

	δ	bias	ESE	SSE	CP
$r = 100$	16%	0.0095	0.3173	0.3153	0.957
	28%	0.0374	0.3551	0.3793	0.950
	38%	0.0436	0.4110	0.4090	0.961
$r = 300$	16%	0.0064	0.1604	0.1555	0.963
	28%	0.0098	0.1808	0.1794	0.955
	38%	0.0011	0.2138	0.2105	0.960
$r = 500$	16%	0.0026	0.1214	0.1201	0.958
	28%	0.0057	0.1360	0.1299	0.970
	38%	0.0051	0.1615	0.1680	0.944

[†] δ is the censoring rate; "bias", "ESE", "SSE" and "CP" are given in the footnotes of Table S.1.

Table S.6

Simulation results on OSP-based $\breve{\theta}_1$ under varying censoring rate (Case IV)†.

	δ	bias	ESE	SSE	CP
$r = 100$	16%	0.0409	0.2334	0.2303	0.959
	28%	0.0465	0.2632	0.2614	0.959
	38%	0.0513	0.2862	0.2786	0.954
$r = 300$	16%	0.0134	0.1253	0.1214	0.961
	28%	0.0233	0.1420	0.1415	0.953
	38%	0.0119	0.1547	0.1490	0.962
$r = 500$	16%	0.0082	0.0954	0.0913	0.966
	28%	0.0097	0.1081	0.1077	0.943
	38%	0.0081	0.1180	0.1134	0.954

† δ is the censoring rate; "bias", "ESE", "SSE" and "CP" are given in the footnotes of Table S.1.

S.2 An additional simulation study

In this section, we conduct a simulation to study the performance of $\breve{\boldsymbol{\theta}}$ if we set $\sum_{i\in S_0}\pi_i = \delta + c$, where $c =-\delta$, -0.2, -0.12, 0, 0.12 and 0.2, respectively. Of note, $c = 0$ leads to $\sum_{i\in S_0}\pi_i = \delta$, which is the adopted setting in our proposed subsampling method. The generation of data is the same as the first simulation study. In Tables S.7 and S.8, we report the bias of OSP-based subsample estimator $\breve{\theta}_1$ (other $\breve{\theta}_i$ are similar) and MSE of $\breve{\boldsymbol{\theta}}$ with $\sum_{i\in S_0}\pi_i = \delta + c$, respectively. More specifically, the subsampling probabilities $\{\pi_i^{m\Gamma}\}_{i=1}^n$ are given by replacing δ with $\delta + c$ in (11). Table S.7 shows that the estimators are unbiased in all cases except for $c = -\delta$. One possible explanation for this phenomenon is due to $\sum_{i\in S_0}\pi_i = 0$ when $c = -\delta$. This case results in the situation that all censored observations cannot be selected into a subsample. Hence, $\mathbf{U}^*(\boldsymbol{\theta})$ in (4) is surely biased towards equation (3) with full data. From the results in Table S.8, $\breve{\boldsymbol{\theta}}$ with $c = 0$ has the smallest MSE. The MSEs become worse as the value of c departs from zero. Hence, it is feasible to set $\sum_{i\in S_0}\pi_i = \delta$ in our method for determining the optimal subsampling probabilities for the AH model.

Table S.7 Bias of $\breve{\theta}_1$ with $\sum_{i \in S_0} \pi_i = \delta + c$.

	Case	$c = -\delta$	$c = -0.2$	$c = -0.12$	$c = 0$	$c = 0.12$	$c = 0.2$
$r = 100$	I	**0.0313**	0.0268	0.0232	0.0199	0.0173	0.0164
	II	**0.0294**	0.0211	0.0244	0.0301	0.0454	0.0440
	III	**0.1208**	0.0216	0.0140	0.0203	0.0214	0.0310
	IV	**0.0143**	0.0823	0.0588	0.0757	0.0677	0.0582
$r = 300$	I	**0.0694**	0.0076	0.0056	0.0133	0.0066	0.0100
	II	**0.0706**	0.0190	0.0018	0.0118	0.0109	0.0097
	III	**0.1368**	0.0010	0.0129	0.0121	0.0021	0.0141
	IV	**0.0293**	0.0390	0.0137	0.0182	0.0193	0.0145
$r = 500$	I	**0.0645**	0.0026	0.0081	0.0041	0.0090	0.0084
	II	**0.0816**	0.0056	0.0029	0.0051	0.0080	0.0020
	III	**0.1243**	0.0097	0.0048	0.0054	0.0018	0.0081
	IV	**0.0787**	0.0224	0.0163	0.0173	0.0125	0.0141

Table S.8 MSE of $\breve{\boldsymbol{\theta}}$ with $\sum_{i \in S_0} \pi_i = \delta + c$.

	Case	$c = -\delta$	$c = -0.2$	$c = -0.12$	$c = 0$	$c = 0.12$	$c = 0.2$
$r = 100$	I	0.5990	0.4673	0.3505	**0.2988**	0.3130	0.3258
	II	0.6316	0.4268	0.3324	**0.2882**	0.2946	0.3149
	III	1.2997	1.5139	1.1011	**0.8563**	0.8812	0.9328
	IV	0.4570	0.3915	0.3129	**0.2872**	0.3042	0.3140
$r = 300$	I	0.1717	0.1325	0.0928	**0.0831**	0.0885	0.0902
	II	0.1753	0.1189	0.0904	**0.0843**	0.0863	0.0871
	III	0.3558	0.4786	0.3018	**0.2311**	0.2368	0.2477
	IV	0.1308	0.1178	0.0864	**0.0758**	0.0845	0.0859
$r = 500$	I	0.1065	0.0742	0.0548	**0.0464**	0.0487	0.0552
	II	0.1132	0.0673	0.0501	**0.0466**	0.0479	0.0514
	III	0.2176	0.2896	0.1707	**0.1263**	0.1279	0.1403
	IV	0.0787	0.0661	0.0487	**0.0446**	0.0469	0.0485