Received: 4 February 2020 | Revised: 30 September 2020 | Accepted: 6 October 2020
DOI: 10.1002/sim.8783

RESEARCH ARTICLE

Statistics in Medicine WILEY

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, China
[2]Department of Statistics, University of Connecticut, Mansfield, Connecticut, USA
[3]Division of Biostatistics, Washington University in St. Louis, St. Louis, Missouri, USA

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

Funding information
Foundation for the National Institutes of Health, Grant/Award Number: NIH UL1 TR002345; National Science Foundation, Grant/Award Number: USA grant DMS-1812013

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, for example, 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 articles 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[15,16] has several advantages: examining additive associations vs multiplicative associations, not assuming PH, 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 article, 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 article 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 article with some discussions. Technical proofs of theoretical results, Tables S.1 to 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, $i=1, \dots, n$. 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 article. The failure indicator is $\Delta_i = I(T_i \le C_i)$, and the censoring rate is $\delta = 1 - n^{-1}\sum_{i=1}^{n}\Delta_i$. Denote the observed-failure counting process by $N_i(t) = I(\tilde{T}_i \le t, \Delta_i = 1)$, and the at-risk indicator by $Y_i(t) = I(\tilde{T}_i \ge t)$. Following Lin and Ying,[16] the intensity of $N_i(t)$ with AH function is

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

where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)'$ is a vector of regression parameters belonging to a compact subset of $\mathbb{R}^p$, $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})'$ 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

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

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

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

where $\psi_i(\boldsymbol{\theta}) = \int_0^{\tau}\{\mathbf{X}_i - \overline{\mathbf{X}}(t)\}\{dN_i(t) - Y_i(t)\boldsymbol{\theta}'\mathbf{X}_i dt\}, i=1, \dots, n$. 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 $i=1, \ldots, r$. 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 $U_i^*(\boldsymbol{\theta}) = \int_0^\tau \{\mathbf{X}_i^* - \overline{\mathbf{X}}^*(t)\}\{dN_i^*(t) - Y_i^*(t)\boldsymbol{\theta}'\mathbf{X}_i^* dt\}$, with $\overline{\mathbf{X}}^*(t) = \{\sum_{i=1}^r \pi_i^{*-1} Y_i^*(t)\mathbf{X}_i^*\}/\{\sum_{i=1}^r \pi_i^{*-1} Y_i^*(t)\}$, $N_i^*(t) = I(\tilde{T}_i^* \leq t, \Delta_i^* = 1)$ and $Y_i^*(t) = I(\tilde{T}_i^* \geq t)$, $i=1, \ldots, r$. 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 $i = 1, \ldots, r$.
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 (OSP) in Section 3. Under Assumptions (A.1) to (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) to (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

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

and

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

3 | SUBSAMPLING STRATEGIES

We consider how to specify the subsampling probabilities $\{\pi_i\}_{i=1}^n$. A naive choice is the uniform subsampling strategy with $\pi_i = n^{-1}$, for $i = 1, \dots, n$. However, these uniform subsampling probabilities (UNIF) may not be optimal, and a nonuniform subsampling method could have a better performance.[7] Our idea is to determine the OSP 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 probabilities by directly minimizing $tr(\boldsymbol{\Gamma})$, which can effectively speed up the subsampling algorithm. Note that

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

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 noncensored individuals. The following result gives the subsampling probabilities $\pi_i^{m\Gamma}$ for $i \in S_1$.

Under Assumptions (A.1) to (A.8) in the Supporting Information, if the subsampling probabilities are chosen as

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

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), $i = 1, \dots, n$. 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

$$\begin{aligned}
\breve{\mathcal{H}} &= \frac{1}{nr}\sum_{i=1}^r \frac{1}{\pi_i^*}\int_0^\tau Y_i^*(t)\{\mathbf{X}_i^* - \overline{\mathbf{X}}^*(t)\}^{\otimes 2} dt, \\
\breve{\boldsymbol{\Gamma}} &= \frac{1}{n^2r^2}\sum_{i=1}^r \frac{1}{\pi_i^{*2}}\int_0^\tau \{\mathbf{X}_i^* - \overline{\mathbf{X}}^*(t)\}^{\otimes 2} dN_i^*(t),
\end{aligned}$$

and $\{\pi_i^*\}_{i=1}^r$ are the corresponding subsampling probabilities for a subsample. The standard errors (SEs) 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 UNIF in comparison to the full data approach, (2) the gain in computation time, and (3) our method's performance with mild vs heavy censoring and how the censoring proportion could affect the choice of r. First, we generate failure times $(T_1, \dots, T_n)$ 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: $\mathbf{X} = (X_1, \dots, X_5)^T$, and X_i are independent exponential random variables with probability density function $f(x) = 2e^{-2x}I(x > 0)$, $i = 1, \dots, 5$.

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 $i = 1, \dots, n$. We carry out computation on a sever with 128 GB memory using R software. In Table 1, we report the estimation results from "the proposed method with OSP" vs "the proposed method with UNIF" for Case I (other cases are given in Tables S.1 to 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

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

where $\breve{\boldsymbol{\theta}}^{(\ell)}$ is from the ℓth replication, $\ell = 1, \dots, L$. 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 different distribution of covariates, we define the estimation efficiency of OSP-based estimator relative to UNIF as

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

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

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,

TABLE 1 Simulation results on the subsample estimator $\breve{\theta}$ with Case I

	r	OSP bias	OSP ESE	OSP SSE	OSP CP	UNIF bias	UNIF ESE	UNIF SSE	UNIF 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

Note: "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 $\hat{\theta}_{ZE}$.

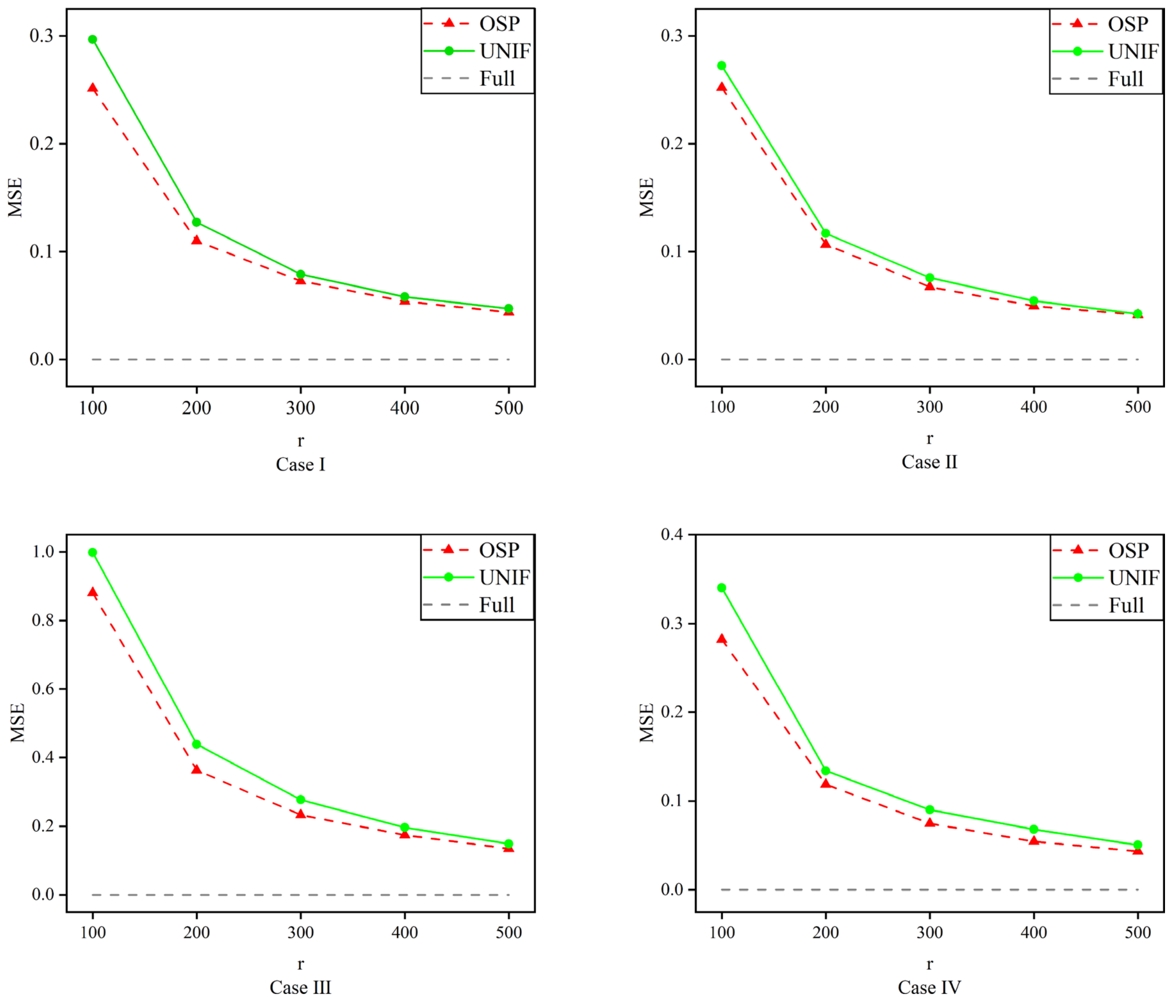

FIGURE 1 The MSEs for different subsampling methods [Color figure can be viewed at wileyonlinelibrary.com]

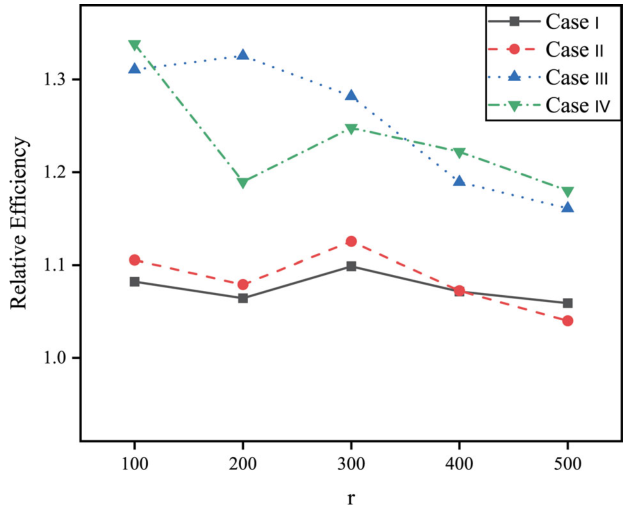

FIGURE 2 Relative efficiency for different settings of covariates [Color figure can be viewed at wileyonlinelibrary.com]

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

Note: "OSP" and "UNIF" are given in the footnotes of Table 1.

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				

Note: "OSP" and "UNIF" are given in the footnotes of Table 1.

for example, 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.

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 to S.6 of the Supporting Information), where $\breve{\theta}_i$ are similar and omitted, for $i = 2, \ldots, 5$. 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.

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

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

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

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)

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

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 subsample estimator (Est), the 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.

6 | CONCLUDING REMARKS

In this article, 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 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

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)

Note: "Bias," "ESE," "SSE," "UNIF," and "OSP" are given in the footnotes of Table 1.

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 (eg, estimating SEs or CIs). To further improve our method, we can consider an iterative subsampling procedure. Specifically, we perform L replications of our proposed approach. Let $\tilde{\theta} = \frac{1}{L}\sum_{\ell=1}^{L}\breve{\theta}^{(\ell)}$, where $\breve{\theta}^{(\ell)}$ is the subsampling-based estimator from the ℓth replication, for $\ell = 1, \ldots, L$. The asymptotic properties of $\tilde{\theta}$ needs further research. Second, the simulations and real data example indicate that the proposed method works well with a moderate subsample size (eg, $r = 500$). 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 AH approach is that the hazard is not constrained to be positive. Therefore, it is interesting to assess the model fit or appropriateness of the AH model in the massive data setting.

ACKNOWLEDGEMENTS

The authors would like to thank the Editor, the Associate Editor and two reviewers for their constructive and insightful comments that greatly improved the article. 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.

DATA AVAILABILITY STATEMENT

We apply our proposed method to a lymphoma cancer dataset in the Surveillance, Epidemiology, and End Results program (https://seer.cancer.gov/).

ORCID

Haixiang Zhang https://orcid.org/0000-0002-7311-5605
Lei Liu https://orcid.org/0000-0003-1844-338X

REFERENCES

1. Zhao T, Cheng G, Liu H. A partially linear framework for massive heterogeneous data. *Ann Stat*. 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. *Ann Stat*. 2018;46(3):1352-1382.
3. Shi C, Lu W, Song R. A massive data framework for M-estimators with cubic-rate. *J Am Stat Assoc*. 2018;113(524):1698-1709.
4. Jordan MI, Lee JD, Yang Y. Communication-efficient distributed statistical inference. *J Am Stat Assoc*. 2019;114(526):668-681.
5. Volgushev S, Chao SK, Cheng G. Distributed inference for quantile regression processes. *Ann Stat*. 2019;47(3):1634-1662.
6. Ma P, Mahoney M, Yu B. A statistical perspective on algorithmic leveraging. *J Mach Learn Res*. 2015;16:861-911.
7. Wang H, Zhu R, Ma P. Optimal subsampling for large sample logistic regression. *J Am Stat Assoc*. 2018;113(522):829-844.
8. Wang H. More efficient estimation for logistic regression with optimal subsample. *J Mach Learn Res*. 2019;20:1-59.
9. Wang H, Yang M, Stufken J. Information-based optimal subdata selection for big data linear regression. *J Am Stat Assoc*. 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. *Comput Stat Data Anal*. 2021;153:107072. https://doi.org/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. https://doi.org/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. *Stat Med*. 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. *J Royal Stat Soc Ser 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). *J Royal Stat Soc Ser 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

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

How to cite this article: Zuo L, Zhang H, Wang HY, Liu L. Sampling-based estimation for massive survival data with additive hazards model. *Statistics in Medicine*. 2020;1–10. https://doi.org/10.1002/sim.8783