Statistica Sinica Preprint No: SS-2022-0028	
Title	Scalable Estimation for High Velocity Survival Data Able to Accommodate Addition of Covariates
Manuscript ID	SS-2022-0028
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0028
Complete List of Authors	Ying Sheng, Yifei Sun, Charles E. McCulloch and Chiung-Yu Huang
Corresponding Authors	Chiung-Yu Huang
E-mails	ChiungYu.Huang@ucsf.edu
Notice: Accepted version subject to English editing.	

SCALABLE ESTIMATION FOR HIGH VELOCITY SURVIVAL DATA ABLE TO ACCOMMODATE ADDITION OF COVARIATES

Ying Sheng[1], Yifei Sun[2], Charles E. McCulloch[3] and Chiung-Yu Huang[3]

[1]*Chinese Academy of Sciences,* [2]*Columbia University and*

[3]*University of California at San Francisco*

Abstract: With the rapidly increasing availability of large-scale streaming data, there has been a growing interest in developing methods that allow the processing of the data in batches without requiring storage of the full dataset. In this paper, we propose a hybrid likelihood approach for scalable estimation of the Cox model using individual-level data in the current data batch and summary statistics calculated from historical data. We show that the proposed scalable estimator is asymptotically as efficient as the maximum likelihood estimator calculated using the entire dataset with low data storage requirements and low loading and computation time. A challenge in analyzing survival data batches that is not accommodated in extant methods is that new covariates may become available midway through data collection. To accommodate addition of covariates, we develop a hybrid empirical likelihood approach to incorporate the historical covariate effects evaluated in a reduced Cox model. The extended scalable estimator is asymptotically more efficient

than the maximum likelihood estimator obtained using only the data batches that include the additional covariates. The proposed approaches are evaluated by numerical simulations and illustrated with an analysis of Surveillance, Epidemiology, and End Results (SEER) breast cancer data.

Key words and phrases: batch processing, hybrid empirical likelihood, scalable estimation

1 Introduction

In recent years, unprecedented technological advances in data collection systems, such as medical devices, health apps, surveillance systems, and wearable sensors, have led to a proliferation of large-scale streaming data. The key characteristics of such data include massive sample size and high velocity, posing challenges in data storage and statistical analysis. As an example, continuous glucose monitors that report blood sugar levels as frequently as once per minute are becoming more common and readily available (Vettoretti et al., 2018). The huge amount of streaming glucose data can provide valuable insight into how well diabetic patients are managing their disease and, in case of frequent hypoglycemia, may promote re-evaluation of care. As another example, the Surveillance, Epidemiology and End Results (SEER) Program began collecting demographic, clinical,

treatment, and outcome variables on nearly all types of incident cancer patients in 1973 (https://seer.cancer.gov). The SEER program has expanded over time to now include 21 cancer registries, covering approximately 35% of the U.S. population. In 2018 alone, data on 724, 852 newly diagnosed cancer cases were submitted to the database, which now has more than 150 data fields. As these data collection systems become widespread, efficient scalable estimation techniques that can process data in batches without high cost of storage are in great need.

Recently, there has been a significant surge in interest in developing approaches that avoid using and storing individual participant data (IPD) for statistical analysis of large-scale data and data batches. For large-scale data, Chen and Xie (2014) proposed a split-and-conquer approach for simultaneous parameter estimation and variable selection. Jordan et al. (2019) presented a surrogate likelihood framework for low-dimensional estimation, high-dimensional regularized estimation, and Bayesian inference that uses distributed computation. For data batches, Schifano et al. (2016) proposed online-updating methods to update parameter estimation sequentially in the linear model and estimating equation framework, while Luo and Song (2020) presented renewable estimation and incremental inference using the current data and historical summary statistics. It is worthwhile pointing

out that many existing meta-analysis approaches, such as Lin and Zeng (2010) and Liu et al. (2015), can also be modified for combining results from different data batches and/or historical information when analyzing large-scale data and/or data batches.

In this paper, we consider the situation where the censored data arrive in sequential batches, and statistical analysis within each batch is feasible. Existing methods for complete data, such as Jordan et al. (2019) and Luo and Song (2020), are only applicable when the likelihood of the full data can be decomposed into a linear combination of the likelihood of each data batch. However, this property does not hold under the Cox model and thus these approaches can not be applied directly to right-censored survival data. To tackle this problem, we propose a hybrid likelihood approach for scalable estimation of the Cox model. Specifically, estimates of unknown parameters can be updated sequentially by synthesizing historical estimates of covariate effects. We show that the proposed scalable estimator can achieve the same estimation efficiency as the oracle maximum likelihood estimator (MLE), which is calculated using the full dataset. Moreover, since only historical summary information and IPD from the current batch are stored at each update, the proposed approach can greatly reduce the data storage cost and is computationally efficient, making it particularly appealing in analyzing

high velocity survival data.

In applications, it is common that new covariates are added during data collection due to advances in scientific knowledge or advances in data collection technology. Since 1988, the SEER program has added more than 40 data fields to the database. For example, after being identified as important predictors of survival outcomes in breast cancer, collaborative stage (CS) tumor size was added to the SEER database in 2004, and human epidermal growth factor receptor 2 (HER2) status was incorporated in 2010. For completely observed data, Wang et al. (2018) proposed a bias-correcting approach by incorporating the cumulative coefficient estimate from reduced working models under generalized linear models. Kundu et al. (2019) developed a generalized meta-analysis approach to estimate the parameter in the maximal model by combining parameter estimates from different reduced models. These approaches, however, are not readily applicable when the observation of the outcome event is subject to right censoring. For large-scale survival data or survival data batches, although researchers have developed estimation procedures utilizing historical estimates or test statistics (see, for example, Xue et al., 2019; Wang et al., 2021; Wu et al., 2021), approaches that apply to accommodate addition of covariates have been lacking.

To handle newly added covariates in survival analysis, we propose a

hybrid empirical likelihood approach (Qin, 2000; Zhang et al., 2020) that exploits summary statistics from historical data with a reduced set of covariates. Although empirical likelihood has gained its popularity in meta-analysis (Chen and Qin, 1993; Qin et al., 2015; Huang et al., 2016; Han and Lawless, 2019), existing approaches are not readily applicable for scalable estimation for high velocity survival data with newly added covariates. We demonstrate that the covariate effects evaluated in a reduced Cox model can be summarized using population estimating equations under the full Cox model. With the available IPD in the current data batch, the population moments can be approximated by the corresponding sample moments. Under the proposed hybrid empirical likelihood framework, the sample estimating equations can be incorporated as constraints to synthesize the historical covariate effect information; moreover, variability in the historical information can be properly accounted for. Compared with the MLE calculated only using batches of IPD that contain the full set of covariates, the proposed scalable estimator enjoys a substantial efficiency gain.

2 Scalable estimation under the Cox model

Let T be the survival time of interest and assume that T is absolutely continuous. Let $\boldsymbol{X}$ denote a d-dimensional vector of baseline covariates.

Denote by $\lambda(t \mid \boldsymbol{X})$ the hazard function of T given $\boldsymbol{X}$. We assume that the survival time T follows the Cox proportional hazards model (Cox, 1972)

$$\lambda(t \mid \boldsymbol{X}) = \lambda(t) \exp(\boldsymbol{\beta}^{\intercal} \boldsymbol{X}), \tag{2.1}$$

where $\boldsymbol{\beta}$ is a d-dimensional vector of regression parameters and $\lambda(t)$ is an unspecified baseline hazard function. Let $\Lambda(t) = \int_0^t \lambda(u)\mathrm{d}u$ be the corresponding baseline cumulative hazard function. Due to end of study or loss to follow-up, the observation of the survival time may be subject to right censoring. Denote by C the censoring time and assume that the survival time T and the censoring time C are independent conditional on $\boldsymbol{X}$. Instead of observing the survival time T, we observe the possibly censored survival time $Y = \min(T, C)$ and the indicator of an observed failure event $\Delta = I(T \leq C)$.

2.1 Estimation of covariate effects

For $b \geq 1$, the bth batch of data consists of n_b independent and identically distributed (i.i.d.) observations. Specifically, denote by Y_{bi}, Δ_{bi}, and $\boldsymbol{X}_{bi}$ the observed survival time, the failure event indicator, and the vector of covariates of the ith observation in the bth batch, $i = 1 \ldots, n_b$. The observed data in the bth batch can be represented using $\mathcal{D}_b = (\boldsymbol{Y}_b, \boldsymbol{\Delta}_b, \boldsymbol{X}_b)$, where

$\boldsymbol{Y}_b = (Y_{b1}, \ldots, Y_{bn_b})$, $\boldsymbol{\Delta}_b = (\Delta_{b1}, \ldots, \Delta_{bn_b})$, and $\boldsymbol{X}_b = (\boldsymbol{X}_{b1}, \ldots, \boldsymbol{X}_{bn_b})$. Note that the bth batch collects the newly added cases only and does not include the patients who are initially collected in previous batches but are still at risk. Assume that the data batches $\mathcal{D}_b$, $b \geq 1$, are independent. Denote by $\mathcal{D}_B^c = \{\mathcal{D}_1, \ldots, \mathcal{D}_B\}$ the cumulative data up to the Bth batch. For $B \geq 1$, let $\widehat{\boldsymbol{\beta}}_B$ be the scalable estimator for $\boldsymbol{\beta}$ up to the Bth data batch. When $B = 1$, the initial scalable estimator $\widehat{\boldsymbol{\beta}}_1$ can be derived as usual by the MLE using IPD in $\mathcal{D}_1$. If we consider an ideal case where $n_1 \to \infty$, it can be shown that $\sqrt{n_1}(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_0)$ converges in distribution to a mean zero multivariate normal random variable with the covariance matrix Σ, where $\boldsymbol{\beta}_0$ is the true value of $\boldsymbol{\beta}$. Moreover, a consistent estimator of Σ, denoted by $\widehat{\Sigma}_1$, can be derived using IPD in $\mathcal{D}_1$. In what follows, a hybrid likelihood approach is presented to update $\widehat{\boldsymbol{\beta}}_B$ using IPD in $\mathcal{D}_B$ and summary statistics $(\widehat{\boldsymbol{\beta}}_{B-1}, \widehat{\Sigma}_{B-1})$ calculated from $\mathcal{D}_{B-1}^c$ for any finite $B \geq 2$.

For $k = 0, 1, 2$, define $S_B^{(k)}(t, \boldsymbol{\beta}) = n_B^{-1} \sum_{i=1}^{n_B} I(Y_{Bi} \geq t) \exp(\boldsymbol{\beta}^\intercal \boldsymbol{X}_{Bi}) \boldsymbol{X}_{Bi}^{\otimes k}$ and let $s^{(k)}(t, \boldsymbol{\beta}) = E\{I(Y \geq t) \exp(\boldsymbol{\beta}^\intercal \boldsymbol{X}) \boldsymbol{X}^{\otimes k}\}$, where $\boldsymbol{x}^{\otimes 0} = 1$, $\boldsymbol{x}^{\otimes 1} = \boldsymbol{x}$ and $\boldsymbol{x}^{\otimes 2} = \boldsymbol{x}\boldsymbol{x}^\intercal$. With the observed IPD in $\mathcal{D}_B$, the log partial likelihood is $\sum_{i=1}^{n_B} \Delta_{Bi} \left[\boldsymbol{\beta}^\intercal \boldsymbol{X}_{Bi} - \log\{S_B^{(0)}(Y_{Bi}, \boldsymbol{\beta})\}\right]$. Following Zhang et al. (2020), we synthesize the historical summary-level information, while accounting for uncertainty therein, from $\mathcal{D}_{B-1}^c$ by treating $\widehat{\boldsymbol{\beta}}_{B-1}$ as the realized value of a

random vector. Generally, $\sqrt{n_{B-1}^c}(\widehat{\boldsymbol{\beta}}_{B-1}-\boldsymbol{\beta}_0)$ converges in distribution to a mean zero multivariate normal random variable with the covariance matrix Σ as $n_{B-1}^c \to \infty$. In the ideal case where Σ is known, we can derive the asymptotic log likelihood, up to a constant, $-n_{B-1}^c(\widehat{\boldsymbol{\beta}}_{B-1}-\boldsymbol{\beta})^\intercal\Sigma^{-1}(\widehat{\boldsymbol{\beta}}_{B-1}-\boldsymbol{\beta})/2$. In practice, the unknown Σ can be replaced with the consistent estimator up to the $(B-1)$th batch, denoted by $\widehat{\Sigma}_{B-1}$, and hence we can derive the hybrid likelihood $\ell_B(\boldsymbol{\beta}) = \sum_{i=1}^{n_B}\Delta_{Bi}\left[\boldsymbol{\beta}^\intercal\boldsymbol{X}_{Bi} - \log\{S_B^{(0)}(Y_{Bi},\boldsymbol{\beta})\}\right] - n_{B-1}^c(\widehat{\boldsymbol{\beta}}_{B-1}-\boldsymbol{\beta})^\intercal\widehat{\Sigma}_{B-1}^{-1}(\widehat{\boldsymbol{\beta}}_{B-1}-\boldsymbol{\beta})/2$. Maximizing the hybrid likelihood $\ell_B(\boldsymbol{\beta})$ yields the scalable estimator, that is,

$$\widehat{\boldsymbol{\beta}}_B = \arg\max_{\boldsymbol{\beta}} \ell_B(\boldsymbol{\beta}). \tag{2.2}$$

Moreover, we propose to update Σ by $\widehat{\Sigma}_B = n_B^c\left(n_{B-1}^c\widehat{\Sigma}_{B-1}^{-1} + n_B\widetilde{\Sigma}_B^{-1}\right)^{-1}$, where $\widetilde{\Sigma}_B$ is a consistent estimator of Σ using the observed IPD in $\mathcal{D}_B$. Given $\widehat{\boldsymbol{\beta}}_B$, the cumulative baseline hazard function can be estimated by the Breslow-type estimator $\widehat{\Lambda}_B(t,\widehat{\boldsymbol{\beta}}_B)$ (Breslow, 1972), where $\widehat{\Lambda}_B(t,\boldsymbol{\beta}) = n_B^{-1}\sum_{i=1}^{n_B}\int_0^t\{S_B^{(0)}(u,\boldsymbol{\beta})\}^{-1}\mathrm{d}N_{Bi}(u)$ with $N_{Bi}(t) = \Delta_{Bi}I(Y_{Bi} \le t)$. The large-sample properties of the proposed scalable estimator $\widehat{\boldsymbol{\beta}}_B$ and $\widehat{\Lambda}_B(t,\widehat{\boldsymbol{\beta}}_B)$ are summarized below in Theorem 1. The derivation of $\widehat{\Sigma}_B$ and the proof of Theorem 1 are given in Section 3 of the Supplementary Materials.

Theorem 1. *Assume* $n_B/n_{B-1}^c \to \kappa_B \in [0,\infty)$ *as* $n_B^c \to \infty$. *Under conditions (C1) and (C2) in the Appendix, as* $n_b \to \infty$, $1 \le b \le B$, *we have (i)* $\sqrt{n_B^c}(\widehat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0)$ *converges in distribution to a mean zero multivariate normal distribution with the covariance matrix* Σ, *where* $\Sigma = \left(\int_0^\infty [s^{(2)}(t,\boldsymbol{\beta}_0) - \{s^{(0)}(t,\boldsymbol{\beta}_0)\}^{-1}\{s^{(1)}(t,\boldsymbol{\beta}_0)\}^{\otimes 2}]\mathrm{d}\Lambda_0(u)\right)^{-1}$; *(ii)* $\sqrt{n_B}\{\widehat{\Lambda}_B(t,\widehat{\boldsymbol{\beta}}_B) - \Lambda_0(t)\}$ *converges in distribution to a mean zero normal distribution with the variance* $\int_0^t \{s^{(0)}(u,\boldsymbol{\beta}_0)\}^{-1}\mathrm{d}\Lambda_0(u) + \kappa_B(1+\kappa_B)^{-1}\mathcal{V}(t)^\intercal \Sigma \mathcal{V}(t)$, *where* $\mathcal{V}(t) = \int_0^t \{s^{(0)}(u,\boldsymbol{\beta}_0)\}^{-1} s^{(1)}(u,\boldsymbol{\beta}_0)\mathrm{d}\Lambda_0(u)$.

Denote by $\widetilde{\boldsymbol{\beta}}_B^c$ the oracle MLE calculated using IPD in the cumulative data $\mathcal{D}_B^c$. It can be shown that as $n_B^c \to \infty$, $\sqrt{n_B^c}(\widetilde{\boldsymbol{\beta}}_B^c - \boldsymbol{\beta}_0)$ converges in distribution to a mean zero multivariate normal distribution with the covariance matrix Σ. Hence by Theorem 1, the proposed scalable estimator $\widehat{\boldsymbol{\beta}}_B$ is asymptotically as efficient as the oracle MLE $\widetilde{\boldsymbol{\beta}}_B^c$. Let $\widetilde{\Lambda}_B(t,\widetilde{\boldsymbol{\beta}}_B) = n_B^{-1}\sum_{i=1}^{n_B}\int_0^t \{S_B^{(0)}(u,\widetilde{\boldsymbol{\beta}}_B)\}^{-1}\mathrm{d}N_{Bi}(u)$, where $\widetilde{\boldsymbol{\beta}}_B$ is the MLE for $\boldsymbol{\beta}$ calculated only using IPD in $\mathcal{D}_B$. It can be shown that $\sqrt{n_B}\{\widetilde{\Lambda}_B(t,\widetilde{\boldsymbol{\beta}}_B) - \Lambda_0(t)\}$ converges in distribution to a mean zero normal distribution with the variance $\int_0^t \{s^{(0)}(u,\boldsymbol{\beta}_0)\}^{-1}\mathrm{d}\Lambda_0(u) + \mathcal{V}(t)^\intercal \Sigma \mathcal{V}(t)$. Hence by incorporating historical covariate effects, the proposed estimator $\widehat{\Lambda}_B(t,\widehat{\boldsymbol{\beta}}_B)$ yields an efficiency gain when compared with $\widetilde{\Lambda}_B(t,\widetilde{\boldsymbol{\beta}}_B)$, which is calculated only using IPD in $\mathcal{D}_B$.

2.2 Equality of covariate effects across batches

The validity of the proposed hybrid likelihood approach holds when the regression coefficient in $\mathcal{D}_B$, denoted by $\boldsymbol{\beta}_B$, is the same as that in $\mathcal{D}^c_{B-1}$, denoted by $\boldsymbol{\beta}^c_{B-1}$. When combining historical summary statistics with current data, there is often a concern that covariate effects may have changed due to changes in disease management or shifts in population demographics. To test the conformity of the historical covariate effects with covariate effects in the current batch, we develop a hybrid likelihood ratio test. We consider the test statistic $R_1 = 2\left\{\sup_{\boldsymbol{\beta}_B,\boldsymbol{\beta}^c_{B-1}} \ell(\boldsymbol{\beta}_B,\boldsymbol{\beta}^c_{B-1}) - \sup_{\boldsymbol{\beta}_B=\boldsymbol{\beta}^c_{B-1}} \ell(\boldsymbol{\beta}_B,\boldsymbol{\beta}^c_{B-1})\right\}$, where $\ell(\boldsymbol{\beta}_B,\boldsymbol{\beta}^c_{B-1}) = \sum_{i=1}^{n_B} \Delta_{Bi}\left[\boldsymbol{\beta}_B^\intercal \boldsymbol{X}_{Bi} - \log\{S_B^{(0)}(Y_{Bi},\boldsymbol{\beta}_B)\}\right] - n^c_{B-1}(\widehat{\boldsymbol{\beta}}_{B-1} - \boldsymbol{\beta}^c_{B-1})^\intercal \widehat{\Sigma}^{-1}_{B-1}(\widehat{\boldsymbol{\beta}}_{B-1} - \boldsymbol{\beta}^c_{B-1})/2$. Note that without the constraint $\boldsymbol{\beta}_B = \boldsymbol{\beta}^c_{B-1}$, the hybrid likelihood $\ell(\boldsymbol{\beta}_B,\boldsymbol{\beta}^c_{B-1})$ is maximized by $\boldsymbol{\beta}_B = \widetilde{\boldsymbol{\beta}}_B$ and $\boldsymbol{\beta}^c_{B-1} = \widehat{\boldsymbol{\beta}}_{B-1}$. Under the conditions specified in Theorem 1 and the null hypothesis $H_0 : \boldsymbol{\beta}_B = \boldsymbol{\beta}^c_{B-1}$, the test statistic R_1 converges in distribution to a χ^2 random variable with d degrees of freedom as $n^c_B \to \infty$. The proof is given in Section 4 of the Supplementary Materials. The hybrid likelihood ratio test is developed to check if a certain batch deviates from the assumed Cox model. In the case where hypotheses arrive sequentially, the $\alpha-$investing method proposed by Foster and Stine (2008) can be used to control the false discovery rate.

3 Scalable estimation incorporating newly added covariates

The premise of the proposed hybrid likelihood approach in Section 2 is that the availability of covariates does not change over time. In practice, however, it is common that new covariates, denoted by $\boldsymbol{W}$, may become available midway through data collection. For example, since 2010, variables including HER2 status and the American Joint Committee on Cancer (AJCC) stage have been added to the SEER database to refine the prognostic information. Since incorporation of the new covariates is likely to improve the prediction accuracy, it is imperative to utilize a model with the full set of covariates. In what follows, we propose to incorporate historical covariate effects evaluated in a reduced model to improve the estimation of the full model.

3.1 Scalable estimation with addition of new covariates

Assume a set of new covariates $\boldsymbol{W}$ is added starting from the B^*th ($B^* > 1$) data batch and let $\boldsymbol{Z} = (\boldsymbol{X}^\intercal, \boldsymbol{W}^\intercal)^\intercal$ be the full set of q covariates ($q > d$). We assume the following Cox model for the survival time T given $\boldsymbol{Z}$,

$$\lambda(t \mid \boldsymbol{Z}) = \lambda^*(t) \exp(\boldsymbol{\theta}^\intercal \boldsymbol{Z}), \tag{3.3}$$

where $\boldsymbol{\theta}$ is a q-dimensional vector of regression parameters and $\lambda^*(t)$ is an unspecified baseline hazard function. Note that the regression parameter $\boldsymbol{\theta}$ in model (3.3) is of interest, although a reduced working model $\lambda(t \mid \boldsymbol{X}) = \lambda(t)\exp(\boldsymbol{\beta}^\intercal \boldsymbol{X})$ is fitted when the new covariates are not available.

The observed IPD in the bth ($b \geq B^*$) data batch are denoted by $\mathcal{D}_b = (\boldsymbol{Y}_b, \boldsymbol{\Delta}_b, \boldsymbol{Z}_b)$. For $k = 0, 1, 2$ and $b \geq B^*$, define functions $S^{(k)}_{b,\boldsymbol{Z}}(t, \boldsymbol{\theta}) = n_b^{-1} \sum_{i=1}^{n_b} I(Y_{bi} \geq t) \exp(\boldsymbol{\theta}^\intercal \boldsymbol{Z}_{bi}) \boldsymbol{Z}_{bi}^{\otimes k}$. Solving the partial score estimating equation $n_{B^*}^{-1} \sum_{i=1}^{n_{B^*}} \int_0^\infty \left\{ \boldsymbol{Z}_{B^*i} - S^{(1)}_{B^*,\boldsymbol{Z}}(Y_{B^*i}, \boldsymbol{\theta}) / S^{(0)}_{B^*,\boldsymbol{Z}}(Y_{B^*i}, \boldsymbol{\theta}) \right\} \mathrm{d}N_{B^*i}(t) = \mathbf{0}$ yields the MLE of $\boldsymbol{\theta}$ based on $\mathcal{D}_{B^*}$, denoted by $\widetilde{\boldsymbol{\theta}}_{B^*}$. The baseline cumulative hazard function $\Lambda^*(t)$ (i.e., $\int_0^t \lambda^*(u)\mathrm{d}u$) can be estimated by the Breslow-type estimator $\widetilde{\Lambda}^*_{B^*}(t, \widetilde{\boldsymbol{\theta}}_{B^*})$, where

$$\widetilde{\Lambda}^*_{B^*}(t, \boldsymbol{\theta}) = \frac{1}{n_{B^*}} \sum_{i=1}^{n_{B^*}} \int_0^t \frac{\mathrm{d}N_{B^*i}(u)}{S^{(0)}_{B^*,\boldsymbol{Z}}(u, \boldsymbol{\theta})}. \tag{3.4}$$

However, the MLE $\widetilde{\theta}_{B^*}$ does not utilize the historical data and thus may not be efficient.

To exploit the historical covariate effects for constructing the hybrid empirical likelihood function, we first establish the asymptotic normality of $\widehat{\boldsymbol{\beta}}_{B^*-1}$, which is the scalable estimator obtained via the proposed hybrid likelihood approach (2.2) in Section 2.1. Although the reduced model is likely to be misspecified, it can be shown that $\sqrt{n^c_{B^*-1}}(\widehat{\boldsymbol{\beta}}_{B^*-1} - \boldsymbol{\beta}_0)$ converges in

distribution to a mean zero normal random variable as $n^c_{B^*-1} \to \infty$. Here the limiting value $\boldsymbol{\beta}_0$ is the unique solution of the following equation,

$$E\left[\int_0^\infty \left\{\boldsymbol{X} - \frac{s^{(1)}(t,\boldsymbol{\beta})}{s^{(0)}(t,\boldsymbol{\beta})}\right\} \mathrm{d}N(t)\right] = \mathbf{0}, \tag{3.5}$$

where $s^{(k)}(t,\boldsymbol{\beta}) = E\{I(Y \geq t)\exp(\boldsymbol{\beta}^\intercal \boldsymbol{X})\boldsymbol{X}^{\otimes k}\}$ for $k = 0, 1$, and the expectations are evaluated under the full Cox model (3.3). Moreover, the asymptotic covariance matrix of $\sqrt{n^c_{B^*-1}}(\widehat{\boldsymbol{\beta}}_{B^*-1} - \boldsymbol{\beta}_0)$ up to the (B^*-1)th data batch can be consistently estimated by $\widehat{\Sigma}_{B^*-1}$, derivation of which is given in Section 5 of the Supplementary Materials.

In what follows, we propose a hybrid empirical likelihood approach to estimate $\boldsymbol{\theta}$ using IPD in $\mathcal{D}_{B^*}$ and summary statistics $(\widehat{\boldsymbol{\beta}}_{B^*-1}, \widehat{\Sigma}_{B^*-1})$. Denote by F the joint distribution function of $(Y, \Delta, \boldsymbol{Z})$. For $i = 1, \ldots, n_{B^*}$, let p_i be the jump of F at $(Y_{B^*i}, \Delta_{B^*i}, \boldsymbol{Z}_{B^*i})$. Then the log empirical likelihood based on the observed data in $\mathcal{D}_{B^*}$ can be expressed as $\sum_{i=1}^{n_{B^*}} \log p_i$. Based on the asymptotic normality of $\widehat{\boldsymbol{\beta}}_{B^*-1}$, we can derive the asymptotic log likelihood $-n^c_{B^*-1}(\widehat{\boldsymbol{\beta}}_{B^*-1} - \boldsymbol{\beta})^\intercal \widehat{\Sigma}^{-1}_{B^*-1}(\widehat{\boldsymbol{\beta}}_{B^*-1} - \boldsymbol{\beta})/2$. Hence the hybrid empirical likelihood is

$$\sum_{i=1}^{n_{B^*}} \log p_i - \frac{n^c_{B^*-1}}{2}(\widehat{\boldsymbol{\beta}}_{B^*-1} - \boldsymbol{\beta})^\intercal \widehat{\Sigma}^{-1}_{B^*-1}(\widehat{\boldsymbol{\beta}}_{B^*-1} - \boldsymbol{\beta}). \tag{3.6}$$

The information from the current data batch $\mathcal{D}_{B^*}$ and the historical cumulative data batches $\mathcal{D}^c_{B^*-1}$ can then be expressed as constraints when maximizing the hybrid empirical likelihood.

First, $\mathcal{D}_{B^*}$ can be used to construct an estimating equation for $\boldsymbol{\theta}_0$ based on the partial score equations under the full Cox model (3.3). Denote by $\Lambda_0^*(t)$ the true value of $\Lambda^*(t)$. For $k = 0, 1$, define functions $s^{(k)}_{\boldsymbol{Z}}(t, \boldsymbol{\theta}) = E\{I(Y \geq t) \exp(\boldsymbol{\theta}^\intercal \boldsymbol{Z}) \boldsymbol{Z}^{\otimes k}\}$. For $i = 1, \ldots, n_{B^*}$, define $g_i(\boldsymbol{\theta}) = \int_0^\infty \{\boldsymbol{Z}_{B^*i} - s^{(1)}_{\boldsymbol{Z}}(t, \boldsymbol{\theta})/s^{(0)}_{\boldsymbol{Z}}(t, \boldsymbol{\theta})\} \{\mathrm{d}N_{B^*i}(t) - I(Y_{B^*i} \geq t) \exp(\boldsymbol{\theta}^\intercal \boldsymbol{Z}_{B^*i}) \mathrm{d}\Lambda_0^*(t)\}$. Note that $\{g_i(\boldsymbol{\theta}), i = 1, \ldots, n_{B^*}\}$ provides the asymptotic i.i.d. representation of the partial score estimating function, or, equivalently, $n_{B^*}^{-1} \sum_{i=1}^{n_{B^*}} \int_0^\infty \{\boldsymbol{Z}_{B^*i} - S^{(1)}_{B^*,\boldsymbol{Z}}(Y_{B^*i}, \boldsymbol{\theta})/S^{(0)}_{B^*,\boldsymbol{Z}}(Y_{B^*i}, \boldsymbol{\theta})\} \mathrm{d}N_{B^*i}(t) = n_{B^*}^{-1} \sum_{i=1}^{n_{B^*}} g_i(\boldsymbol{\theta}) + o_p(n_{B^*}^{-1/2})$. Moreover, it can be shown that $E\{g_i(\boldsymbol{\theta}_0)\} = \mathbf{0}$. Heuristically, a set of constraints can be constructed as $\sum_{i=1}^{n_{B^*}} p_i g_i(\boldsymbol{\theta}) = \mathbf{0}$. Using $\mathcal{D}_{B^*}$, we replace the unknown functions $s^{(k)}_{\boldsymbol{Z}}(\cdot, \boldsymbol{\theta})$ with their empirical estimates $S^{(k)}_{B^*,\boldsymbol{Z}}(\cdot, \boldsymbol{\theta})$ and replace $\Lambda_0^*(\cdot)$ with the Breslow-type estimator $\widetilde{\Lambda}^*_{B^*}(\cdot, \boldsymbol{\theta})$ defined by (3.4). Hence we can construct the constraint $\sum_{i=1}^{n_{B^*}} p_i \widehat{g}_i(\boldsymbol{\theta}) = \mathbf{0}$, where $\widehat{g}_i(\boldsymbol{\theta}) = \int_0^\infty \{\boldsymbol{Z}_{B^*i} - S^{(1)}_{B^*,\boldsymbol{Z}}(t, \boldsymbol{\theta})/S^{(0)}_{B^*,\boldsymbol{Z}}(t, \boldsymbol{\theta})\} \{\mathrm{d}N_{B^*i}(t) - I(Y_{B^*i} \geq t) \exp(\boldsymbol{\theta}^\intercal \boldsymbol{Z}_{B^*i}) \mathrm{d}\widetilde{\Lambda}^*_{B^*}(t, \boldsymbol{\theta})\}$.

Second, motivated by Equation (3.5) from the reduced working model, we can further construct another estimating equation for $(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0)$. Define $h_i(\boldsymbol{\theta}, \boldsymbol{\beta}) = \int_0^\infty \{\boldsymbol{X}_{B^*i} - s^{(1)}(t, \boldsymbol{\beta})/s^{(0)}(t, \boldsymbol{\beta})\} \{\mathrm{d}N_{B^*i}(t) - I(Y_{B^*i} \geq t) \exp(\boldsymbol{\beta}^\intercal \boldsymbol{X}_{B^*i})$

$s_{\boldsymbol{Z}}^{(0)}(t,\boldsymbol{\theta})/s^{(0)}(t,\boldsymbol{\beta})\mathrm{d}\Lambda_0^*(t)\}$. It can be shown that $E\{h_i(\boldsymbol{\theta}_0,\boldsymbol{\beta}_0)\} = \mathbf{0}$ and $n_{B^*}^{-1}\sum_{i=1}^{n_{B^*}}\int_0^\infty\{\boldsymbol{X}_{B^*i}-S_{B^*}^{(1)}(Y_{B^*i},\boldsymbol{\beta})/S_{B^*}^{(0)}(Y_{B^*i},\boldsymbol{\beta})\}\mathrm{d}N_{B^*i}(t) = n_{B^*}^{-1}\sum_{i=1}^{n_{B^*}}h_i(\boldsymbol{\theta},\boldsymbol{\beta})+o_p(n_{B^*}^{-1/2})$. By replacing the unknown functions with the corresponding empirical estimates using observed IPD in $\mathcal{D}_{B^*}$, we construct the constraint $\sum_{i=1}^{n_{B^*}}p_i\widehat{h}_i(\boldsymbol{\beta},\boldsymbol{\theta}) = \mathbf{0}$, where $\widehat{h}_i(\boldsymbol{\beta},\boldsymbol{\theta}) = \int_0^\infty\{\boldsymbol{X}_{B^*i}-S_{B^*}^{(1)}(t,\boldsymbol{\beta})/S_{B^*}^{(0)}(t,\boldsymbol{\beta})\}\{\mathrm{d}N_{B^*i}(t)-I(Y_{B^*i}\geq t)\exp(\boldsymbol{\beta}^\intercal\boldsymbol{X}_{B^*i})S_{B^*,\boldsymbol{Z}}^{(0)}(t,\boldsymbol{\theta})/S_{B^*}^{(0)}(t,\boldsymbol{\beta})\mathrm{d}\widetilde{\Lambda}_{B^*}^*(t,\boldsymbol{\theta})\}$.

We propose to maximize the hybrid empirical likelihood in (3.6) subject to the constraints

$$p_i \geq 0,\ \sum_{i=1}^{n_{B^*}}p_i = 1,\ \sum_{i=1}^{n_{B^*}}p_i\widehat{g}_i(\boldsymbol{\theta}) = \mathbf{0},\ \sum_{i=1}^{n_{B^*}}p_i\widehat{h}_i(\boldsymbol{\theta},\boldsymbol{\beta}) = \mathbf{0}.$$

The Lagrange function for the constrained maximization problem is $\mathcal{L} = \sum_{i=1}^{n_{B^*}}\log p_i - n_{B^*-1}^c(\widehat{\boldsymbol{\beta}}_{B^*-1}-\boldsymbol{\beta})^\intercal\widehat{\Sigma}_{B^*-1}^{-1}(\widehat{\boldsymbol{\beta}}_{B^*-1}-\boldsymbol{\beta})/2 - n_B\xi_0\,(\sum_{i=1}^{n_{B^*}}p_i-1) - n_{B^*}\sum_{i=1}^{n_{B^*}}p_i\boldsymbol{\xi}_1^\intercal\widehat{g}_i(\boldsymbol{\theta}) - n_{B^*}\sum_{i=1}^{n_{B^*}}p_i\boldsymbol{\xi}_2^\intercal\widehat{h}_i(\boldsymbol{\theta},\boldsymbol{\beta})$, where ξ_0, $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ are Lagrange multipliers. Taking the derivative of the objective function $\mathcal{L}$ with respect to p_i and setting the derivative equal to 0 yields $\xi_0 = 1$ and $\widehat{p}_i = n_{B^*}^{-1}\{1+\boldsymbol{\xi}_1^\intercal\widehat{g}_i(\boldsymbol{\theta})+\boldsymbol{\xi}_2^\intercal\widehat{h}_i(\boldsymbol{\theta},\boldsymbol{\beta})\}^{-1}$. Moreover, the Lagrange multipliers $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ are determined by $n_{B^*}^{-1}\sum_{i=1}^{n_{B^*}}\{1+\boldsymbol{\xi}_1^\intercal\widehat{g}_i(\boldsymbol{\theta})+\boldsymbol{\xi}_2^\intercal\widehat{h}_i(\boldsymbol{\theta},\boldsymbol{\beta})\}^{-1}\widehat{g}_i(\boldsymbol{\theta}) = \mathbf{0}$ and $n_{B^*}^{-1}\sum_{i=1}^{n_{B^*}}\{1+\boldsymbol{\xi}_1^\intercal\widehat{g}_i(\boldsymbol{\theta})+\boldsymbol{\xi}_2^\intercal\widehat{h}_i(\boldsymbol{\theta},\boldsymbol{\beta})\}^{-1}\widehat{h}_i(\boldsymbol{\theta},\boldsymbol{\beta}) = \mathbf{0}$. Substituting $\widehat{p}_i$ back to the Lagrange function yields the constrained hybrid empirical likelihood

$\min_{\boldsymbol{\xi}_1,\boldsymbol{\xi}_2} \ell(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\xi}_1,\boldsymbol{\xi}_2)$, where, up to a constant,hybrid likelihood approach.

$$\begin{aligned}\ell(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\xi}_1,\boldsymbol{\xi}_2) &= -\sum_{i=1}^{n_{B^*}} \log\left\{1+\boldsymbol{\xi}_1^\intercal \widehat{g}_i(\boldsymbol{\theta})+\boldsymbol{\xi}_2^\intercal \widehat{h}_i(\boldsymbol{\theta},\boldsymbol{\beta})\right\} \\ &\quad -\frac{n^c_{B^*-1}}{2}(\widehat{\boldsymbol{\beta}}_{B^*-1}-\boldsymbol{\beta})^\intercal \widehat{\Sigma}^{-1}_{B^*-1}(\widehat{\boldsymbol{\beta}}_{B^*-1}-\boldsymbol{\beta}).\end{aligned} \tag{3.7}$$

Arguing as in Newey and Smith (2004), the proposed constrained maximization can be carried out by solving the following optimization problem,

$$(\widehat{\boldsymbol{\beta}}_{B^*},\widehat{\boldsymbol{\theta}}_{B^*}) = \arg\max_{\boldsymbol{\beta},\boldsymbol{\theta}} \min_{\boldsymbol{\xi}_1,\boldsymbol{\xi}_2} \ell(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\xi}_1,\boldsymbol{\xi}_2). \tag{3.8}$$

The large-sample properties of the proposed scalable estimator $\widehat{\boldsymbol{\theta}}_{B^*}$ are summarized below in Theorem 2, with the proof given in Section 6 of the Supplementary Materials.

Theorem 2. *Assume that $n_{B^*}/n^c_{B^*-1} \to \kappa_{B^*}$ as $n^c_{B^*} \to \infty$. Under conditions (C3)-(C5) in the Appendix, as $n_b \to \infty$, $1 \le b \le B^*$, $\sqrt{n_{B^*}}(\widehat{\boldsymbol{\theta}}_{B^*}-\boldsymbol{\theta}_0)$ converges in distribution to a zero mean multivariate normal distribution with the covariance matrix $V\{V^{-1}-(1+\kappa_{B^*})^{-1}\Omega^\intercal H\Omega\}V$, where V, Ω and H are given in the Supplementary Materials.*

In the proof of Theorem 2, we show that $\sqrt{n_B}(\widetilde{\boldsymbol{\theta}}_{B^*}-\boldsymbol{\theta}_0)$ converges in distribution to a zero mean multivariate normal distribution with the covariance matrix V as $n_{B^*} \to \infty$. Hence the proposed scalable estimator

$\widehat{\boldsymbol{\theta}}_{B^*}$ is asymptotically more efficient than $\widetilde{\boldsymbol{\theta}}_{B^*}$, with larger efficiency gains for smaller values of κ_{B^*}. In the case with equal sized data batches, the asymptotic covariance matrix of $\widehat{\boldsymbol{\theta}}_{B^*}$ is $V\{V^{-1}-(B^*-1)\Omega^\intercal H\Omega/B^*\}V$ and thus the efficiency gain increases with B^*. When B^* is large, the covariance matrix is approximately $V(V^{-1}-\Omega^\intercal H\Omega)V$ and $\widehat{\boldsymbol{\theta}}_{B^*}$ enjoys a substantial efficiency gain.

Moreover, based on Theorem 2, we can estimate the asymptotic covariance matrix by $\widehat{\Pi}_{B^*}(\widehat{\boldsymbol{\theta}}_{B^*},\widehat{\boldsymbol{\beta}}_{B^*})$, where

$$\widehat{\Pi}_{B^*}(\boldsymbol{\theta},\boldsymbol{\beta})=\widehat{V}(\boldsymbol{\theta})\left\{\widehat{V}(\boldsymbol{\theta})^{-1}-c_{B^*}\widehat{\Omega}(\boldsymbol{\theta},\boldsymbol{\beta})\widehat{H}(\boldsymbol{\theta},\boldsymbol{\beta})\widehat{\Omega}(\boldsymbol{\theta},\boldsymbol{\beta})^\intercal\right\}\widehat{V}(\boldsymbol{\theta}), \quad (3.9)$$

with $\widehat{V}(\boldsymbol{\theta})$, $\widehat{H}(\boldsymbol{\theta},\boldsymbol{\beta})$, and $\widehat{\Omega}(\boldsymbol{\theta},\boldsymbol{\beta})$ given in Section 6 of the Supplementary Materials and $c_{B^*}=\left(1+n_{B^*}/n^c_{B^*-1}\right)^{-1}$.

3.2 Scalable estimation with the full set of covariates

We then apply the hybrid likelihood approach in Section 2.1 to update $\widehat{\boldsymbol{\theta}}_B$, $B\geq B^*+1$. We note that the convergence rate of $\widehat{\boldsymbol{\theta}}_{B-1}$ is controlled by the cumulative sample size of data batches with fully observed $\boldsymbol{Z}$, that is, $n^c_{B-1}-n^c_{B^*-1}$, because the new covariates $\boldsymbol{W}$ are not observed in the first B^*-1 data batches. It can be shown that as $n_b\to\infty$, $1\leq b\leq B-1$, $\sqrt{n^c_{B-1}-n^c_{B^*-1}}(\widehat{\boldsymbol{\theta}}_{B-1}-\boldsymbol{\theta}_0)$ converges in distribution to a mean zero

multivariate normal distribution and the asymptotic covariance matrix can be consistently estimated by $\widehat{\Pi}_{B-1}$ up to the $(B-1)$th batch. Applying the hybrid likelihood approach in Section 2.1, $\boldsymbol{\theta}$ can be estimated by

$$\widehat{\boldsymbol{\theta}}_B = \arg\max_{\boldsymbol{\theta}} \left\{ \ell_B(\boldsymbol{\theta}) - \frac{n^c_{B-1} - n^c_{B^*-1}}{2} (\widehat{\boldsymbol{\theta}}_{B-1} - \boldsymbol{\theta})^\intercal \widehat{\Pi}^{-1}_{B-1} (\widehat{\boldsymbol{\theta}}_{B-1} - \boldsymbol{\theta}) \right\}, \quad (3.10)$$

where $\ell_B(\boldsymbol{\theta}) = \sum_{i=1}^{n_B} \Delta_{Bi} [\boldsymbol{\theta}^\intercal \boldsymbol{Z}_{Bi} - \log\{S^{(0)}_{B,\boldsymbol{Z}}(Y_{Bi}, \boldsymbol{\theta})\}]$ is the partial likelihood based on the IPD in $\mathcal{D}_B$. Moreover, we can update the asymptotic covariance matrix by

$$\widehat{\Pi}_B = (n^c_B - n^c_{B^*-1}) \{ n_B \widehat{V}_B(\widehat{\boldsymbol{\theta}}_B)^{-1} + (n^c_{B-1} - n^c_{B^*-1}) \widehat{\Pi}^{-1}_{B-1} \}^{-1}. \quad (3.11)$$

The large-sample properties of $\widehat{\boldsymbol{\theta}}_B$ are summarized in Theorem 3. The derivation of $\widehat{\Pi}_B$ and the proof of Theorem 3 are given in Section 7 of the Supplementary Materials.

Theorem 3. *For $B > B^*$, assume that $n_{B^*}(n^c_B - n^c_{B^*-1})^{-1} \to r_B \in [0, \infty)$ as $n^c_B \to \infty$. Under the conditions specified in Theorem 2, as $n_b \to \infty$, $1 \le b \le B$, $\sqrt{n^c_B - n^c_{B^*-1}}(\widehat{\boldsymbol{\theta}}_B - \boldsymbol{\theta}_0)$ converges in distribution to a mean zero multivariate normal distribution with the covariance matrix $(V^{-1} + r_B \Omega \widetilde{H} \Omega^\intercal)^{-1}$, where V, Ω and $\widetilde{H}$ are given in the Supplementary Materials.*

Denote by $\widetilde{\boldsymbol{\theta}}^*_B$ the MLE calculated using IPD in $\{\mathcal{D}_{B^*}, \ldots, \mathcal{D}_B\}$, that is,

the data batches with fully observed $\boldsymbol{Z}$. We have $\sqrt{n_B^c - n_{B^*-1}^c}(\widetilde{\boldsymbol{\theta}}_B^* - \boldsymbol{\theta}_0)$ converges in distribution to a mean zero multivariate normal distribution with the covariance matrix V as $n_B^c \to \infty$,. Hence the proposed estimator $\widehat{\boldsymbol{\theta}}_B$ enjoys an efficiency gain when compared with the MLE $\widetilde{\boldsymbol{\theta}}_B^*$. In the case with equal size data batches, we have $r_B = (B - B^* + 1)^{-1}$, where $B - B^* + 1$ is the number of data batches with fully observed $\boldsymbol{Z}$. When $B = B^* + 1$, we have $r_B = 1/2$ and the efficiency gain is substantial. Moreover, when the number of data batches with fully observed $\boldsymbol{Z}$ is very large, we have $r_B \to 0$ and not unexpectedly, the efficiency gain is limited.

3.3 Conformity of the covariate effects information

The validity of the proposed hybrid empirical likelihood approach in Section 3.1 holds when the historical covariate effect information is consistent with the current IPD, which is equivalent to the null hypothesis $H_0 : \boldsymbol{\xi}_2 = \mathbf{0}$. Motivated by Qin and Lawless (1994) and Qin and Lawless (1995), we develop a hybrid empirical likelihood ratio test to check the conformity of the historical covariate effects information in a reduced model. To be specific, we consider the test statistic $R_2 = 2\left\{\sup_{\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\xi}_1,\boldsymbol{\xi}_2} \ell(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\xi}_1,\boldsymbol{\xi}_2) - \sup_{\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\xi}_1} \ell(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\xi}_1,\mathbf{0})\right\}$, where $\ell(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\xi}_1,\boldsymbol{\xi}_2)$ is defined by (3.7). Under the conditions specified in Theorem 2 and the null hypothesis $H_0 : \boldsymbol{\xi}_2 = \mathbf{0}$, the test statistic R_2 con-

verges in distribution to a χ^2 random variable with d degrees of freedom as $n_{B^*} \to \infty$. The proof of the large-sample properties of the test statistic R_2 is given in Section 8 of the Supplementary Materials.

4 Computation

In this section, we focus on the computational aspects of the algorithms. The outline of the algorithm used for scalable estimation when new covariates are added (i.e., Section 3) is described in Section 4.1. Scalable estimation based on the hybrid likelihood in Section 2 can be implemented using the Newton-Raphson method and thus is omitted. We then give some details on comparison of computational efficiency.

4.1 Algorithm

Step 0. Initialization: Obtain initial values $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\Sigma}_1$ using $\mathcal{D}_1$.

Step 1. Scalable estimation under the reduced Cox model: For $B = 2, \ldots, B^* - 1$, calculate $\widehat{\boldsymbol{\beta}}_B$ and $\widehat{\Sigma}_B$ by maximizing the hybrid likelihood using the Newton-Raphson method. The details are given in Section 5 of the Supplementary Materials.

Step 2. Scalable estimation with addition of new covariates: For $B = B^*$, calculate $\widehat{\boldsymbol{\theta}}_{B^*}$ and by solving the constrained maximization in (3.8) and calculate $\widehat{\Pi}_{B^*}$ using Equation (3.9). This can be solved using a nested coordinate descent algorithm presented below.

Step 3. Scalable estimation with the full set of covariates: For $B \geq B^*+1$, calculate $\widehat{\boldsymbol{\theta}}_B$ by maximizing the hybrid likelihood in (3.10) using Newton-Raphson method and calculate $\widehat{\Pi}_B$ using Equation (3.11).

In what follows, a nested coordinate descent algorithm is developed to solve the constrained maximization $\max_{\boldsymbol{\beta},\boldsymbol{\theta}} \min_{\boldsymbol{\xi}_1,\boldsymbol{\xi}_2} \ell(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2)$ in Step 2. When such min-max representation is available, the nested optimization algorithm has been commonly adopted to obtain the empirical likelihood estimator (Chen et al., 2002; Imbens, 2002; Han and Lawless, 2019). Define $\ell(\boldsymbol{\gamma}, \boldsymbol{\xi}) = \ell(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2)$ and $\ell(\boldsymbol{\gamma}) = \min_{\boldsymbol{\xi}} \ell(\boldsymbol{\gamma}, \boldsymbol{\xi})$, where $\boldsymbol{\gamma} = (\boldsymbol{\theta}^\intercal, \boldsymbol{\beta}^\intercal)^\intercal$ and $\boldsymbol{\xi} = (\boldsymbol{\xi}_1^\intercal, \boldsymbol{\xi}_2^\intercal)^\intercal$. For each $\boldsymbol{\gamma}$, $\ell(\boldsymbol{\gamma}, \boldsymbol{\xi})$ is a strictly convex function of $\boldsymbol{\xi}$, and thus $\ell(\boldsymbol{\gamma})$ can be easily evaluated. Moreover, $\ell(\boldsymbol{\gamma})$ is an asymptotically concave function of $\boldsymbol{\gamma}$ in the sense that $\partial^2 \ell(\boldsymbol{\gamma})/\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^\intercal = \mathcal{J}_\gamma(\boldsymbol{\gamma}, \boldsymbol{\xi}(\boldsymbol{\gamma})) + o_p(1)$, where $\mathcal{J}_\gamma(\boldsymbol{\gamma}, \boldsymbol{\xi}(\boldsymbol{\gamma}))$ is a negative definite matrix. The details are given in Section 9 of the Supplementary Materials. Given the above properties of $\ell(\boldsymbol{\gamma}, \boldsymbol{\xi})$, we propose a nested optimization algorithm consisting of two loops: the outer loop maximizes $\ell(\boldsymbol{\gamma})$ using the Newton method, and the inner loop

calculates $\ell(\boldsymbol{\gamma})$ by minimizing $\ell(\boldsymbol{\gamma}, \boldsymbol{\xi})$ for each given value of $\boldsymbol{\gamma}$. In both loops, solving the $(d+q)$-dimensional optimization problems with respect to $\boldsymbol{\xi}$ or $\boldsymbol{\gamma}$ could be computationally challenging due to the complicated forms of the Hessian matrices. Motivated by the coordinate descent algorithm (Wright, 2015), we solve the inner loop optimization by updating $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ sequentially and solve the outer loop optimization by updating $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ sequentially. In this way, each subproblem is a lower-dimensional problem, and thus can be solved more easily than the original problems. The nested coordinate descent algorithm is described below.

Outer Loop

Step 0: Set $l = 0$ and obtain initial values $\boldsymbol{\gamma}^{(0)}=(\boldsymbol{\theta}^{(0)\intercal}, \boldsymbol{\beta}^{(0)\intercal})^\intercal = (\widetilde{\boldsymbol{\theta}}_{B^*}^\intercal, \widehat{\boldsymbol{\beta}}_{B^*-1}^\intercal)^\intercal$, where $\widetilde{\boldsymbol{\theta}}_{B^*}$ is the MLE calculated using $\mathcal{D}_{B^*}$.

Step $(l+1)$: At the $(l+1)$th iteration, calculate $\boldsymbol{\theta}^{(l+1)} = \boldsymbol{\theta}^{(l)} - \tau_1\{\mathcal{J}_{\boldsymbol{\theta}}(\boldsymbol{\theta}^{(l)}, \boldsymbol{\beta}^{(l)}, \boldsymbol{\xi}(\boldsymbol{\gamma}^{(l)}))\}^{-1}\mathcal{U}_{\boldsymbol{\theta}}(\boldsymbol{\theta}^{(l)}, \boldsymbol{\beta}^{(l)}, \boldsymbol{\xi}(\boldsymbol{\gamma}^{(l)}))$ and $\boldsymbol{\beta}^{(l+1)} = \boldsymbol{\beta}^{(l)} - \tau_1\{\mathcal{J}_{\boldsymbol{\beta}}(\boldsymbol{\theta}^{(l+1)}, \boldsymbol{\beta}^{(l)}, \boldsymbol{\xi}(\boldsymbol{\gamma}^{(l)})\}^{-1}\mathcal{U}_{\boldsymbol{\beta}}(\boldsymbol{\theta}^{(l+1)}, \boldsymbol{\beta}^{(l)}, \boldsymbol{\xi}(\boldsymbol{\gamma}^{(l)})$, where Jacobian matrices $\mathcal{U}_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\xi}(\boldsymbol{\gamma}))$ and $\mathcal{U}_{\boldsymbol{\beta}}(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\xi}(\boldsymbol{\gamma}))$, and Hessian matrices $\mathcal{J}_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\xi}(\boldsymbol{\gamma}))$ and $\mathcal{J}_{\boldsymbol{\beta}}(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\xi}(\boldsymbol{\gamma}))$ are given in Section 9 of the Supplementary Materials, and τ_1 is a step-length parameter to avoid overshooting; set $\boldsymbol{\gamma}^{(l+1)} = (\boldsymbol{\theta}^{(l+1)\intercal}, \boldsymbol{\beta}^{(l+1)\intercal})^\intercal$.

Repeat step $(l+1)$ until $\|\boldsymbol{\gamma}^{(l+1)} - \boldsymbol{\gamma}^{(l)}\|$ is smaller than a pre-

specified threshold.

Inner Loop

Step 0: Set $k = 0$ and obtain initial values $\boldsymbol{\xi}_1^{(k)} = \boldsymbol{\xi}_2^{(k)} = \mathbf{0}$; let $\boldsymbol{\xi}^{(k)} = (\boldsymbol{\xi}_1^{(k)\intercal}, \boldsymbol{\xi}_2^{(k)\intercal})^\intercal$;

Step $(k+1)$: At the $(k+1)$th iteration, given $\boldsymbol{\gamma}$, calculate $\boldsymbol{\xi}_1^{(k+1)} = \boldsymbol{\xi}_1^{(k)} - \tau_2\{\mathcal{J}_{\boldsymbol{\xi}_1}(\boldsymbol{\gamma}, \boldsymbol{\xi}_1^{(k)}, \boldsymbol{\xi}_2^{(k)})\}^{-1}\mathcal{U}_{\boldsymbol{\xi}_1}(\boldsymbol{\gamma}, \boldsymbol{\xi}_1^{(k)}, \boldsymbol{\xi}_2^{(k)})$ and $\boldsymbol{\xi}_2^{(k+1)} = \boldsymbol{\xi}_2^{(k)} - \tau_2\{\mathcal{J}_{\boldsymbol{\xi}_2}(\boldsymbol{\gamma}, \boldsymbol{\xi}_1^{(k+1)}, \boldsymbol{\xi}_2^{(k)})\}^{-1}\mathcal{U}_{\boldsymbol{\xi}_2}(\boldsymbol{\gamma}, \boldsymbol{\xi}_1^{(k+1)}, \boldsymbol{\xi}_2^{(k)})$ where Jacobian matrices $\mathcal{U}_{\boldsymbol{\xi}_1}(\boldsymbol{\gamma}, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2)$ and $\mathcal{U}_{\boldsymbol{\xi}_2}(\boldsymbol{\gamma}, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2)$, and Hessian matrices $\mathcal{J}_{\boldsymbol{\xi}_1}(\boldsymbol{\gamma}, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2)$ and $\mathcal{J}_{\boldsymbol{\xi}_2}(\boldsymbol{\gamma}, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2)$ are given in Section 9 of the Supplementary Materials, and τ_2 is a step-length parameter; let $\boldsymbol{\xi}^{(k+1)}(\boldsymbol{\gamma}) = (\boldsymbol{\xi}_1^{(k+1)\intercal}, \boldsymbol{\xi}_2^{(k+1)\intercal})^\intercal$. Repeat step $(k+1)$ until $\|\boldsymbol{\xi}^{(k+1)}(\boldsymbol{\gamma}) - \boldsymbol{\xi}^{(k)}(\boldsymbol{\gamma})\|$ is smaller than a pre-specified threshold.

In the inner loop, the objective functions are strictly convex functions of $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ and thus the iterative algorithm of the optimization almost always converges (Chen et al., 2002; Han and Lawless, 2019). In the outer loop, the convergence of the proposed algorithm is usually fast since initial values $\widetilde{\boldsymbol{\theta}}_{B^*}$ and $\widehat{\boldsymbol{\beta}}_{B^*-1}$ are consistent estimators of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, respectively. Following Han and Lawless (2019), one can sequentially try step-lengths $1, 2^{-1}, \ldots, 2^{-5}$ for updating the unknown parameters. The first length that

makes the value of objective function increase is accepted, and the step length 2^{-5} is used if no such step-length is found.

4.2 Computational efficiency

We conducted Monte-Carlo simulations to compare computational efficiency and required memory of the proposed based scalable estimator in Section 2.1 and the oracle MLE calculated using the entire dataset. Details of the simulations are given in Section 1 of the Supplementary Materials. The computation was performed using the `R` statistical software version 3.6.0 on UCSF Computation Biology and Informatics core at the shared high-performance computing cluster C4 with 150GB of memory. The oracle MLE was calculated using the function `coxph` in the `R` package `survival`. The evaluation criteria for computational efficiency included (i) the total time spent in loading the data; (ii) the total computation time, which refers to the total amount of time required by data loading and algorithm execution, and (iii) the total storage memory required.

As shown in Table 1 in the Supplementary Materials, the proposed approach can greatly reduce the computation time for total computation time and, not surprisingly, gains more computational efficiency with larger cumulative sample size n_B^c and the dimensionality of covariates p. The

ratios of computation time of the proposed scalable estimator to that of the oracle MLE ranged from 2.8 to 4.1 across different values of n_B^c and p. Also, in the case of $n_B^c = 10^8$ and $p = 50$, it was infeasible to calculate the oracle MLE using the entire dataset while the proposed approach was able to complete the computation in a few hours.

5 Simulations and data analysis

5.1 Numerical simulations

We conducted Monte-Carlo simulations to examine the finite-sample performance of the proposed methods. In all simulations, 1,000 datasets, each with 100,000 observations, were generated and then divided into 100 data batches with equal sample sizes. The censoring time C was generated from uniform distributions that yielded censoring rates of 25%, 50% and 75%. In our simulations, the numerical performance of the nested coordinate descent algorithm in Section 4 was not sensitive to the selection of step lengths and thus we set $\tau_1 = \tau_2 = 1$ to save computation time.

In the first set of simulations, we investigated the performance of the proposed hybrid likelihood approach (see Section 2.1) for scalable estimation under the Cox model. The covariates X_1, X_2, X_3 were independently generated from the standard normal distribution, X_4 was gener-

ated from a Bernoulli distribution with $\Pr(X_4 = 1) = 0.5$, and X_5 was generated from a Bernoulli distribution with $\Pr(X_5 = 1) = 0.25$. The survival time T was generated from the Cox proportional hazards model $\lambda(t \mid \boldsymbol{X}) = 1.5\sqrt{t}\exp(\boldsymbol{\beta}^\intercal \boldsymbol{X})$, where $\boldsymbol{\beta} = (0.5, 0.5, 0.5, -1, -1)^\intercal$. Table 1 summarizes the simulation results of the oracle MLE $\widetilde{\boldsymbol{\beta}}_B^c$ calculated using all 100,000 observations, the proposed scalable estimator $\widehat{\boldsymbol{\beta}}_B$, the inverse-variance estimator $\widetilde{\boldsymbol{\beta}}_B^{LZ} = (\sum_{b=1}^B \widehat{\Sigma}_b^{-1})^{-1} \sum_{b=1}^B \widehat{\Sigma}_b^{-1} \widehat{\boldsymbol{\beta}}_b$ proposed in Lin and Zeng (2010), the estimator $\widetilde{\boldsymbol{\beta}}_B^{APBC}$ calculated by the adaptive partition and bias correction (APBC) method proposed in Wu et al. (2021), and the MLE $\widetilde{\boldsymbol{\beta}}_B$ calculated only using 1,000 observations in the last batch $\mathcal{D}_B$. The estimator $\widetilde{\boldsymbol{\beta}}_B^{APBC}$ can be calculated using the `R` package `updatesurvival` by setting the initial number of intervals as $J_0 = 5$. As shown in Table 1, the performance of the scalable estimator $\widehat{\boldsymbol{\beta}}_B$ is similar to that of the oracle MLE $\widetilde{\boldsymbol{\beta}}_B^c$, the inverse-variance estimator $\widetilde{\boldsymbol{\beta}}_B^{LZ}$, and the estimator $\widetilde{\boldsymbol{\beta}}^{APBC}$. We also estimated the cumulative baseline hazard function at 500 time points, which were equally spaced between 0 and 1. The proposed estimator $\widehat{\Lambda}_B(t)$ enjoys an efficiency gain when compared with the MLE $\widetilde{\Lambda}_B(t)$ which was calculated only using 1,000 observations in the last batch $\mathcal{D}_B$, with a relative efficiency ranging from 1.17 to 1.63 across selected time points.

In the second set of simulations, we investigated the performance of

Table 1: Summary of simulation results under the Cox model

		$\widetilde{\boldsymbol{\beta}}_B^c$			$\widehat{\boldsymbol{\beta}}_B$			$\widetilde{\boldsymbol{\beta}}_B^{LZ}$			$\widetilde{\boldsymbol{\beta}}_B^{APBC}$			$\widetilde{\boldsymbol{\beta}}_B$		
Cen		Bias	SE	SEE	Bias	SE	SEE	Bias	SE	SEE	Bias	SE	SEE	Bias	SE	SEE
25%	β_1	0	40	39	-2	40	39	9	40	40	-1	40	39	43	412	396
	β_2	0	39	39	-1	39	39	9	39	40	-2	39	39	43	401	395
	β_3	-1	39	39	-2	39	39	8	39	40	-3	39	39	22	397	396
	β_4	-1	78	78	5	78	78	-6	78	78	6	78	78	-66	819	784
	β_5	2	96	93	8	96	94	-8	97	94	7	96	95	-30	921	942
50%	β_1	1	47	47	-1	47	47	13	48	48	-1	47	47	53	482	477
	β_2	0	47	47	-2	47	47	13	48	48	-1	47	47	48	478	477
	β_3	-1	48	47	-3	48	47	11	48	48	-3	49	47	19	494	478
	β_4	2	93	94	8	93	95	-6	94	95	7	93	95	-99	970	953
	β_5	3	122	120	10	122	121	-16	123	121	10	122	120	5	1161	1213
75%	β_1	2	67	66	-1	67	66	11	68	67	1	67	66	100	667	671
	β_2	0	65	66	-3	65	66	19	66	67	3	66	66	49	659	670
	β_3	-1	65	66	-4	66	66	18	67	67	-3	67	66	53	698	671
	β_4	-4	140	137	20	140	137	-25	142	138	19	141	138	-114	1397	1381
	β_5	6	190	185	30	190	185	-14	194	186	28	190	186	-66	1936	1871

NOTE: Cen, the censoring rate; the true values of the regression coefficients $(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$ are $(0.5, 0.5, 0.5, -1, -1)$; $\widetilde{\boldsymbol{\beta}}_B^c$, the oracle maximum likelihood estimator calculated using all of the data; $\widehat{\boldsymbol{\beta}}_B$, the proposed hybrid likelihood scalable estimator; $\widetilde{\boldsymbol{\beta}}_B^{LZ}$, the inverse-variance estimator proposed in Lin and Zeng (2010); $\widetilde{\boldsymbol{\beta}}_B^{APBC}$, the estimator calculated by the adaptive partition and bias correction (APBC) method proposed in Wu et al. (2021); $\widetilde{\boldsymbol{\beta}}_B$, the maximum likelihood estimator calculated using the data in the Bth data batch; Bias, SE and SEE, empirical bias ($\times 10^4$), empirical standard error ($\times 10^4$) and empirical mean of the standard error estimates ($\times 10^4$).

the proposed hybrid empirical likelihood approach (see Section 3) for scalable estimation when new covariates are added. We generated covariates $\boldsymbol{X} = (X_1, X_2)^\intercal$ from a mean zero bivariate normal random vector with $\text{var}(X_1) = 1, \text{var}(X_2) = 1$, and a correlation coefficient of 0.5. Two new covariates W_1 and W_2 were added since the 51th data batch, where W_1 was generated from a logistic regression model with $\Pr(W_1 = 1 \mid X_1) = \exp(0.5X_1)/\{1 + \exp(0.5X_1)\}$ and $W_2 = W_1X_1$. This specification aims to mimic the situation where a new treatment indicator, which is correlated with currently observed covariates, and its interaction with a covariate are added to the Cox model. The survival time T was generated from the maximal Cox model $\lambda(t \mid \boldsymbol{Z}) = 1.5\sqrt{t}\exp(\boldsymbol{\theta}^\intercal\boldsymbol{Z})$, where $\boldsymbol{Z} = (X_1, X_2, W_1, W_2)^\intercal$ and $\boldsymbol{\theta} = (0.5, 0.5, -1, -1)^\intercal$. For the first 50 data batches, only a reduced set of covariates $\boldsymbol{X}$ was observed and the full set of covariates $\boldsymbol{Z}$ was observed starting from the 51th data batch, so $B^* = 51$. When applying the proposed hybrid empirical approach to estimate the regression coefficient, we incorporated the updated estimates of the covariate effects of $\boldsymbol{X}$ in a reduced Cox model. The simulation results are presented in Table 2. As shown in the table, the proposed scalable estimator $\widehat{\boldsymbol{\theta}}_B$ outperforms the MLE $\widetilde{\boldsymbol{\theta}}_B$, which is calculated only using IPD from the current batch $\mathcal{D}_B$, with smaller standard errors. Moreover, the proposed estimator $\widehat{\boldsymbol{\theta}}_B$ enjoys

an efficiency gain when compared with the MLE $\widetilde{\boldsymbol{\theta}}_B^*$, which is calculated using IPD from batches with newly added covariates, that is, $\{\mathcal{D}_{B^*}, \ldots, \mathcal{D}_B\}$. When the sample size of the historical data is much larger than that of data with newly added covariates, that is, $B = 51$, the proposed estimator $\widehat{\boldsymbol{\theta}}_B$ enjoys a substantial efficiency gain over the MLE $\widetilde{\boldsymbol{\theta}}_B^*$ calculated using IPD from the 51th batch. The relative efficiency ranges from 1.20 to 3.81 in estimating $\boldsymbol{\beta}$ across different censoring rates. As expected, the efficiency gain decreases as the sample size of data with newly added covariates increases. In the case where the sample size of the historical data and that of data with newly added covariates are comparable, that is, $B = 100$, the relative efficiency ranges from 1.04 to 1.29, when compared with the MLE $\widetilde{\boldsymbol{\theta}}_B^*$ calculated using IPD from batches 51–100.

5.2 Analysis of SEER breast cancer dataset

Breast cancer is the most frequently diagnosed cancer among women in the United States. Accurate prediction of the breast cancer mortality risk is essential for successful breast cancer management. The Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute (NCI) collects demographics and cancer factors on all types of incident cancer patients from cancer registries covering over 30% of the U.S. population. Such a large-scale dataset provides valuable resources to

Table 2: Summary of simulation results in the presence of new covariates

		$\widehat{\boldsymbol{\theta}}_B$			$\widetilde{\boldsymbol{\theta}}^*_B$			$\widetilde{\boldsymbol{\theta}}_B$		
Cen		Bias	SE	SEE	Bias	SE	SEE	Bias	SE	SEE
$B = 51$										
25%	θ_1	5	495	429	5	595	584	5	595	584
	θ_2	3	255	225	22	458	449	22	458	449
	θ_3	-59	707	661	-58	818	789	-58	818	789
	θ_4	9	778	726	-11	853	814	-11	853	814
50%	θ_1	28	563	506	15	706	687	15	706	687
	θ_2	5	286	269	21	549	541	21	549	541
	θ_3	-24	871	832	-60	981	948	-60	981	948
	θ_4	36	956	883	-20	1052	998	-20	1052	998
75%	θ_1	23	669	604	24	954	922	24	954	922
	θ_2	14	404	365	19	787	758	19	787	758
	θ_3	-89	1226	1174	-98	1416	1389	-98	1416	1389
	θ_4	6	1339	1276	4	1499	1456	4	1499	1456
$B = 100$										
25%	θ_1	-1	76	77	-1	85	82	9	592	584
	θ_2	2	58	59	4	65	63	59	462	450
	θ_3	-10	105	106	-5	108	111	-32	763	790
	θ_4	-2	107	108	-4	115	114	-47	828	813
50%	θ_1	-4	90	91	-1	97	96	17	682	687
	θ_2	4	69	71	5	79	76	57	546	542
	θ_3	-7	127	130	-6	133	130	-29	947	949
	θ_4	-3	132	133	-3	145	140	-76	1006	998
75%	θ_1	-3	124	120	-3	132	129	-2	916	923
	θ_2	8	96	98	6	108	106	82	764	761
	θ_3	-9	190	185	-7	196	194	-49	1384	1390
	θ_4	7	199	192	3	207	203	-39	1059	1453

NOTE: Cen, the censoring rate; the true values of the regression coefficients $(\theta_1, \theta_2, \theta_3, \theta_4)$ are $(0.5, 0.5, -1, -1)$; the full set of covariates are observed starting from the B^*th data batch with $B^* = 51$; $\widehat{\boldsymbol{\theta}}_B$, the proposed hybrid empirical likelihood scalable estimator; $\widetilde{\boldsymbol{\theta}}^*_B$, the maximum likelihood estimator calculated using data batches that include the additional covariates; $\widetilde{\boldsymbol{\theta}}_B$, the maximum likelihood estimator calculated using the data in the Bth data batch; Bias, SE and SEE, empirical bias ($\times 10^4$), empirical standard error ($\times 10^4$) and empirical mean of the standard error estimates ($\times 10^4$).

evaluate the effects of patients' demographics and tumor characteristics on survival. The breast cancer data from SEER have the characteristics of large sample size and high velocity. As time goes by, the updates in the SEER database include not only newly diagnosed breast cancer cases, but also new variables for improved characterization of patients' risk profile. In what follows, we applied the proposed scalable approaches to build risk prediction models for breast cancer mortality.

Our study sample consisted of patients who were diagnosed with breast cancer during 2000-2010 and the event of interest is death due to breast cancer. The covariates collected since 2000 included age at diagnosis (≥ 50 years vs. < 50 years), estrogen receptors (ER, positive vs. negative), progesterone receptors (PR, positive vs. negative), cancer grade (III/IV vs. I/II), race (White, African American, and other). In response to the demand of having better prognosis and predictive factors in evaluating and guiding breast cancer treatment, the SEER Program began collecting new data items related to breast cancer prognosis under the Collaborative Stage (CS) Data Collection System since 2004 and an updated edition (CSv2) since 2010 (Howlader et al., 2014). Hence data on CS tumor size are available only for cases diagnosed after 2004, while HER2 and AJCC stage 7th edition are available only for cases diagnosed after 2010.

Excluding the patients with missing covariate values, the dataset consists of 192,404 observations with a total of 9,822 (5.1%) events being observed during the follow-up period. We divided the dataset into 11 data batches according to the year of diagnosis, where the Bth batch includes breast cancer cases diagnosed in year $1999 + B$. The covariate CS tumor size is available starting with the 5th batch and covariates HER2 and AJCC stage are available starting with the 11th batch. Our goal is to build a prediction model that utilize these important biomarkers while combining historical covariate effect estimates using batches that do not include these newly added predictors. To this end, we applied the hybrid empirical likelihood approach in Section 3 to update the risk prediction models.

Table 3 summarizes the point estimates and standard errors of the proposed scalable estimator $\widehat{\boldsymbol{\theta}}_B$ and the MLE $\widetilde{\boldsymbol{\theta}}_B^*$, and $\widetilde{\boldsymbol{\theta}}_B^*$ was calculated using IPD in the batches which share the same set of covariates. As shown in Table 3, the two methods yield similar coefficient estimates. More importantly, the proposed scalable estimator $\widehat{\boldsymbol{\theta}}_B$ enjoys efficiency gains when compared with the MLE $\widetilde{\boldsymbol{\theta}}_B^*$ by incorporating the historical covariate effect information. Using cases diagnosed during 2000–2003, the proposed scalable estimator $\widehat{\boldsymbol{\theta}}_B$ yields similar results compared to the MLE $\widetilde{\boldsymbol{\theta}}_B^*$ calculated using IPD from 2000–2003. Using cases diagnosed up to 2009, the proposed

estimator $\widehat{\boldsymbol{\theta}}_B$ enjoys an efficiency gain when compared with the MLE $\widetilde{\boldsymbol{\theta}}_B^*$ calculated using IPD from 2004 to 2009. The relative efficiency ranges from 1.02 to 1.20. At year 2010, the proposed estimator $\widehat{\boldsymbol{\theta}}_B$ enjoys an efficiency gain when compared with the MLE $\widetilde{\boldsymbol{\theta}}_B^*$ calculated using IPD in 2010 and the relative efficiency ranges from 1.02 to 6.83. The substantial efficiency gain of the proposed scalable approach is attributed to the information from the historical data, whose sample size ($n = 174,899$) is much larger than the batch size in 2010 ($n = 17,505$). Moreover, the effect of CS tumor size (2–4 cm vs < 2 cm) did not reach statistical significance when MLE was applied but was statistically significant when the proposed scalable approach was applied. Adjusting for other covariates, a larger tumor size (2-4 cm vs < 2 cm) is associated with a shorter length of breast cancer survival, with hazard ratio of $\exp(0.39) \approx 1.48$ (95% CI, 1.22-1.77).

6 Discussion

In this article, we proposed a hybrid empirical likelihood framework for scalable estimation with survival data batches. Our estimation procedure is flexible in that it can incorporate various forms of summary statistics from historical batches. Moreover, the proposed approach can greatly reduce the storage memory and computation time for data loading and algorithm ex-

Table 3: Estimated regression coefficients of the Cox model for the breast cancer study

	$B=4$		$B=10$		$B=11$	
	$\widehat{\boldsymbol{\theta}}_B$	$\widetilde{\boldsymbol{\theta}}_B^*$	$\widehat{\boldsymbol{\theta}}_B$	$\widetilde{\boldsymbol{\theta}}_B^*$	$\widehat{\boldsymbol{\theta}}_B$	$\widetilde{\boldsymbol{\theta}}_B^*$
age ($\geq$ 50 years)	0.239	0.240	0.432	0.448	0.496	0.439
	(0.038)	(0.038)	(0.032)	(0.035)	(0.036)	(0.095)
ER negative	0.623	0.623	0.403	0.389	0.508	0.510
	(0.049)	(0.049)	(0.037)	(0.041)	(0.049)	(0.108)
PR negative	0.538	0.538	0.595	0.602	0.739	0.819
	(0.049)	(0.049)	(0.037)	(0.041)	(0.040)	(0.108)
grade (III&IV vs. I&II)	0.970	0.971	0.534	0.533	0.496	0.501
	(0.040)	(0.040)	(0.032)	(0.034)	(0.037)	(0.092)
race (white vs others)	0.132	0.141	0.308	0.361	0.537	0.557
	(0.066)	(0.066)	(0.052)	(0.055)	(0.066)	(0.163)
race (black vs others)	0.666	0.670	0.641	0.703	0.838	0.827
	(0.075)	(0.075)	(0.058)	(0.061)	(0.077)	(0.178)
CS tumor size (2-4 cm vs <2cm)	-	-	1.114	1.117	0.385	0.241
	-	-	(0.037)	(0.038)	(0.096)	(0.141)
CS tumor size ($\geq$ 4 cm vs <2cm)	-	-	2.112	2.116	0.746	0.801
	-	-	(0.037)	(0.038)	(0.102)	(0.138)
HER2 negative	-	-	-	-	0.559	0.503
	-	-	-	-	(0.101)	(0.102)
AJCC stage I vs 0	-	-	-	-	1.476	1.309
	-	-	-	-	(0.919)	(1.008)
AJCC stage II vs 0	-	-	-	-	2.161	2.138
	-	-	-	-	(0.924)	(1.005)
AJCC stage III vs 0	-	-	-	-	3.259	3.256
	-	-	-	-	(0.926)	(1.005)
AJCC stage IV vs 0	-	-	-	-	4.797	4.723
	-	-	-	-	(0.927)	(1.005)

NOTE: ER, estrogen receptors; PR, progesterone receptors; CS tumor size, Collaborative Stage tumor size; HER2, human epidermal growth factor receptor 2; AJCC stage, American Joint Committee for Cancer stage grouping, 7th edition; the entire dataset is divided into 11 data batches according to the year of diagnosis, where the Bth batch includes breast cancer cases diagnosed in year $1999 + B$. The covariate CS tumor size is available starting with the 5th batch and covariates HER2 and AJCC stage are available starting with the 11th batch; $\widehat{\boldsymbol{\theta}}_B$, the proposed hybrid empirical likelihood scalable estimator; $\widetilde{\boldsymbol{\theta}}_B^*$, the maximum likelihood estimator calculated using the data batches which share the same set of covariates; standard error estimates are given in the parentheses.

ecution when compared to standard estimation. Our approach can also accommodate the addition of covariates over time and can achieve significant efficiency gains compared to using only the batches of data with complete covariate information. We developed two testing procedures to check the homogeneity assumption for the proposed scalable estimation methods. Rejection of the null hypothesis indicates potential violation of the assumption and hence scalable estimation using the proposed approaches may not be reliable. When the homogeneity assumption is violated, one needs to postulate proper statistical models to accommodate the batch heterogeneity and extend the scalable estimation methods to account for such heterogeneity. This will be studied in our future research.

Supplementary Materials

The Supplementary Materials contain additional numerical simulations, additional analysis of SEER breast cancer dataset, the proofs of Theorems 1–3, and details of the proposed computation algorithm.

Appendix

We adopt the following regularity conditions:

(C1) The vector of covariates $\boldsymbol{X}$ is bounded with probability one. The

true value $\boldsymbol{\beta}_0$ lies in a compact subset of $\mathbb{R}^d$.

(C2) The censoring time C is conditionally independent of T given $\boldsymbol{X}$.

(C3) The vector of covariates $\boldsymbol{Z}$ is bounded with probability one. The true value $\boldsymbol{\theta}_0$ lies in a compact subset of $\mathbb{R}^q$.

(C4) The censoring time C is conditionally independent of T given $\boldsymbol{Z}$.

(C5) Let $\boldsymbol{\gamma} = (\boldsymbol{\theta}^\intercal, \boldsymbol{\beta}^\intercal)^\intercal$ and let $\boldsymbol{\gamma}_0 = (\boldsymbol{\theta}_0^\intercal, \boldsymbol{\beta}_0^\intercal)^\intercal$ be the true value of $\boldsymbol{\gamma}$. For $k = 0, 1$, define $s_{\boldsymbol{Z}}^{(k)}(t, \boldsymbol{\theta}) = E\{I(Y \geq t) \exp(\boldsymbol{\theta}^\intercal \boldsymbol{Z}) \boldsymbol{Z}^k\}$ and $s^{(k)}(t, \boldsymbol{\beta}) = E\{I(Y \geq t) \exp(\boldsymbol{\beta}^\intercal \boldsymbol{X}) \boldsymbol{X}^k\}$. Let $U(\boldsymbol{\gamma}) = (g(\boldsymbol{\theta})^\intercal, h(\boldsymbol{\beta}, \boldsymbol{\theta})^\intercal)^\intercal$, where

$$g(\boldsymbol{\theta}) = \int_0^\infty \left\{ \boldsymbol{Z} - \frac{s_{\boldsymbol{Z}}^{(1)}(t, \boldsymbol{\theta})}{s_{\boldsymbol{Z}}^{(0)}(t, \boldsymbol{\theta})} \right\} \{ \mathrm{d}N(t) - I(Y \geq t) \exp(\boldsymbol{\theta}^\intercal \boldsymbol{Z}) \mathrm{d}\Lambda_0^*(t) \},$$

$$h(\boldsymbol{\beta}, \boldsymbol{\theta}) = \int_0^\infty \left\{ \boldsymbol{X} - \frac{s^{(1)}(t, \boldsymbol{\beta})}{s^{(0)}(t, \boldsymbol{\beta})} \right\} \left\{ \mathrm{d}N(t) - I(Y \geq t) \exp(\boldsymbol{\beta}^\intercal \boldsymbol{X}) \frac{s_{\boldsymbol{Z}}^{(0)}(t, \boldsymbol{\theta})}{s^{(0)}(t, \boldsymbol{\beta})} \mathrm{d}\Lambda_0^*(t) \right\}.$$

The functions $\partial U(\boldsymbol{\gamma})/\partial \boldsymbol{\gamma}$ and $\partial^2 U(\boldsymbol{\gamma})/\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^\intercal$ are continuous in a neighborhood of $\boldsymbol{\gamma}_0$. Moreover, functions $\|U(\boldsymbol{\gamma})\|^3$, $\|\partial U(\boldsymbol{\gamma})/\partial \boldsymbol{\gamma}\|$ and $\|\partial^2 U(\boldsymbol{\gamma})/\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^\intercal\|$ are bounded by some integrable functions in this neighborhood.

References

Breslow, N. E. (1972). Discussion of the paper by D. R. Cox. *Journal of the Royal Statistical Society, Series B 34*, 216–217.

Chen, J. and J. Qin (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika 80*(1), 107–116.

Chen, J., R. Sitter, and C. Wu (2002). Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika 89*(1), 230–237.

Chen, X. and M.-g. Xie (2014). A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica 24*, 1655–1684.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B 34*(2), 187–202.

Foster, D. P. and R. A. Stine (2008). α-investing: a procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society: Series B 70*(2), 429–444.

Han, P. and J. F. Lawless (2019). Empirical likelihood estimation using auxiliary summary information with different covariate distributions. *Statistica Sinica 29*, 1321–1342.

Howlader, N., V. W. Chen, L. A. Ries, M. M. Loch, R. Lee, C. DeSantis, C. C. Lin, J. Ruhl, and K. A. Cronin (2014). Overview of breast cancer collaborative stage data items – their definitions, quality, usage, and

clinical implications: a review of seer data for 2004-2010. *Cancer 120*, 3771–3780.

Huang, C.-Y., J. Qin, and H.-T. Tsai (2016). Efficient estimation of the Cox model with auxiliary subgroup survival information. *Journal of the American Statistical Association 111*(514), 787–799.

Imbens, G. W. (2002). Generalized method of moments and empirical likelihood. *Journal of Business & Economic Statistics 20*(4), 493–506.

Jordan, M. I., J. D. Lee, and Y. Yang (2019). Communication-efficient distributed statistical inference. *Journal of the American Statistical Association 114*(526), 668–681.

Kundu, P., R. Tang, and N. Chatterjee (2019). Generalized meta-analysis for multiple regression models across studies with disparate covariate information. *Biometrika 106*(3), 567–585.

Lin, D. and D. Zeng (2010). On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. *Biometrika 97*(2), 321–332.

Liu, D., R. Y. Liu, and M.-g. Xie (2015). Multivariate meta-analysis of heterogeneous studies using only summary statistics: efficiency and robustness. *Journal of the American Statistical Association 110*(509), 326–340.

Luo, L. and P. X.-K. Song (2020). Renewable estimation and incremental inference in generalized linear models with streaming data sets. *Journal of the Royal Statistical Society: Series B 82*(1), 69–97.

Newey, W. K. and R. J. Smith (2004). Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica 72*(1), 219–255.

Qin, J. (2000). Combining parametric and empirical likelihoods. *Biometrika 87*(2), 484–490.

Qin, J. and J. Lawless (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics 22*(1), 300–325.

Qin, J. and J. Lawless (1995). Estimating equations, empirical likelihood and constraints on parameters. *Canadian Journal of Statistics 23*(2), 145–159.

Qin, J., H. Zhang, P. Li, D. Albanes, and K. Yu (2015). Using covariate-specific disease prevalence information to increase the power of case-control studies. *Biometrika 102*(1), 169–180.

Schifano, E. D., J. Wu, C. Wang, J. Yan, and M.-H. Chen (2016). Online updating of statistical inference in the big data setting. *Technometrics 58*(3), 393–403.

Vettoretti, M., G. Cappon, G. Acciaroli, A. Facchinetti, and G. Sparacino (2018). Continuous glucose monitoring: current use in diabetes management and possible future applications. *Journal of Diabetes Science and Technology 12*(5), 1064–1071.

Wang, C., M.-H. Chen, J. Wu, J. Yan, Y. Zhang, and E. Schifano (2018). Online updating method with new variables for big data streams. *Canadian Journal of Statistics 46*(1), 123–146.

Wang, Y., C. Hong, N. Palmer, Q. Di, J. Schwartz, I. Kohane, and T. Cai (2021). A fast divide-and-conquer sparse Cox regression. *Biostatistics 22*(2), 381–401.

Wright, S. J. (2015). Coordinate descent algorithms. *Mathematical Programming 151*(1), 3–34.

Wu, J., M.-H. Chen, E. D. Schifano, and J. Yan (2021). Online updating of survival analysis. *Journal of Computational and Graphical Statistics 30*(4), 1209–1223.

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

Zhang, H., L. Deng, M. Schiffman, J. Qin, and K. Yu (2020). Generalized

integration model for improved statistical inference by leveraging external summary data. *Biometrika 107*, 689–703.

Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, 100190, China

E-mail: shengying@amss.ac.cn

Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY 10032, U.S.A.

E-mail: ys3072@cumc.columbia.edu

Department of Epidemiology & Biostatistics, University of California at San Francisco, San Francisco , CA 94158, U.S.A.

E-mail: Charles.McCulloch@ucsf.edu

Department of Epidemiology & Biostatistics, University of California at San Francisco, San Francisco , CA 94158, U.S.A.

E-mail: ChiungYu.Huang@ucsf.edu (corresponding author)