Biometrics 000, 000–000 DOI: 000
000 0000

An Online Updating Approach for Testing the Proportional Hazards Assumption with Streams of Survival Data

Yishu Xue*, HaiYing Wang, Jun Yan***, and Elizabeth D. Schifano******

Department of Statistics, University of Connecticut, Storrs, Connecticut, U.S.A.

**email:* yishu.xue@uconn.edu

***email:* haiying.wang@uconn.edu

****email:* jun.yan@uconn.edu

*****email:* elizabeth.schifano@uconn.edu

Summary: The Cox model, which remains as the first choice in analyzing time-to-event data even for large datasets, relies on the proportional hazards (PH) assumption. When survival data arrive sequentially in chunks, a fast and minimally storage intensive approach to test the PH assumption is desirable. We propose an online updating approach that updates the standard test statistic as each new block of data becomes available, and greatly lightens the computational burden. Under the null hypothesis of PH, the proposed statistic is shown to have the same asymptotic distribution as the standard version computed on the entire data stream with the data blocks pooled into one dataset. In simulation studies, the test and its variant based on most recent data blocks maintain their sizes when the PH assumption holds and have substantial power to detect different violations of the PH assumption. We also show in simulation that our approach can be used successfully with "big data" that exceed a single computer's computational resources. The approach is illustrated with the survival analysis of patients with lymphoma cancer from the Surveillance, Epidemiology, and End Results Program. The proposed test promptly identified deviation from the PH assumption that was not captured by the test based on the entire data.

Key words: Cox model; diagnostics; Schoenfeld residuals.

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

Recent advances in information technology have made available data that arrive in high velocity everyday. Online methods, such as the online updating estimation and inference presented in Schifano et al. (2016), are appealing as storage of historical data is not required which yields great savings in computing resources. Survival data, or time-to-event data, may also arrive sequentially, and the desire for online updated inferences in the survival setting is not uncommon. For example, flight information, such as delay time until take-off or cancellation, is available for more than 114,000 commercial flights scheduled daily around the world (Air Transport Action Group, 2018); real estate information, such as time on market until sold, is updated continuously for the over 6 million homes in the real-estate market (National Association of Realtors, 2018). As such events occur everyday at high frequency, observations also accumulate quickly.

The Cox model (Cox, 1972) is the most commonly used tool in analyzing survival data. A crucial step in fitting the popular Cox model is to check the proportional hazards (PH) assumption (e.g., Xue and Schifano, 2017). The standard approach, if new data becomes available along a stream, would be to pool all historical data together, fit a new Cox model, and use standard methods such as the test of Grambsch and Therneau (1994) to examine whether the PH assumption is appropriate. This, however, can pose a heavy computational burden and can be very time-consuming when the data size gets large. While efforts have been made in fitting Cox model using distributed computing and therefore reducing the computing time, such as in Wang et al. (2019), methods for checking the PH assumption in these settings have not been developed.

In this work, we propose a method to test the PH assumption in the online updating setting, which does not require storage or access to the historical data. Our approach is an application of the divide-and-conquer and online updating strategies (Lin and Xi, 2011;

Schifano et al., 2016) to the streaming survival data setting. The data is assumed to arrive sequentially in blocks, an the test statistic is an appropriately aggregated version of the standard test statistic of Grambsch and Therneau (1994) computed from each block. The statistics can be adapted to be based on data in a moving window of certain size, which may be more useful in detecting local deviations from the null hypothesis. A byproduct of our method is a cumulatively updated estimating equation (CUEE) estimator for the regression coefficients if the PH assumption is not rejected.

When the null hypothesis of PH is true, our test statistic is shown to have the same asymptotic distribution as the standard (full data) statistic under certain regularity conditions. In simulation studies, under the null hypothesis, the proposed test holds its size and the CUEE estimator closely approximates the estimator based on the full data; when the null hypothesis is not true, the test has comparable or higher power than the standard statistic based on the full data. For a dataset that can be loaded into computer memory, our proposed statistic can be computed in significantly less time than the standard statistic. Our test can also successfully be used within a reasonable amount of time for big data that cannot (easily) be loaded into memory. The method is illustrated by analyzing the survival time of the lymphoma cancer patients in the Surveillance, Epidemiology, and End Results (SEER) Program. Interestingly, while the changes in parameters were not captured by using the standard (full data) test of Grambsch and Therneau (1994), they were promptly identified by our online updated version.

The rest of this article is organized as follows. In Section 2, we review the notation of the Cox model and the test statistic of Grambsch and Therneau (1994). In Section 3, we propose our online updating test statistics for the PH assumption. We present simulation results in Section 4, and illustrate the usage of the test with an application to the survival time of patients with lymphoma cancer from the SEER data in Section 5. A discussion concludes

in Section 6. The proposed methods are all implemented in R based on functions from the **survival** package (Therneau, 2015), and the code can be found via GitHub (Xue, 2018).

2. Cox Proportional Hazards Model

2.1 *Notation and Preliminaries*

For completeness we review the Cox model and tests for the PH assumption. Let T_i^* be the true event time and C_i be the censoring time for subject i. Define $T_i = \min(T_i^*, C_i)$ and $\delta_i = I(T_i^* \leqslant C_i)$. Suppose we observe independent copies of $(\delta_i, T_i, \boldsymbol{X}_i)$, $i = 1, \ldots, n$, where $\boldsymbol{X}_i$ is the p-dimensional vector of covariates of the ith subject. The Cox model specifies the hazard for individual i as

$$\lambda_i(t) = \lambda_0(t) \exp\left(\boldsymbol{X}_i^\top \boldsymbol{\beta}\right), \tag{1}$$

where λ_0 is an unspecified non-negative function of time called the baseline hazard, and $\boldsymbol{\beta}$ is a p-dimensional coefficient vector in a compact parameter space. Because the logarithm of the hazard ratio for two subjects with fixed covariate vectors $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$, $(\boldsymbol{X}_i - \boldsymbol{X}_j)^\top \boldsymbol{\beta}$, is proportional to the difference in covariate values and is otherwise constant over time ($\boldsymbol{\beta}$), the model is also known as the PH model. It has been later extended to incorporate time-dependent covariates. For the rest of the article, we use $\boldsymbol{X}_i(t)$ to indicate the possibility of covariates being time-dependent.

Cox (1972, 1975) formulated the partial likelihood approach to estimate $\boldsymbol{\beta}$. For untied failure time data, Fleming and Harrington (1991) expressed it under the counting process formulation to be

$$\mathrm{PL}(\boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{t \geqslant 0} \left[\frac{Y_i(t) \exp\left\{\boldsymbol{X}_i(t)^\top \boldsymbol{\beta}\right\}}{\sum_j Y_j(t) \exp\left\{\boldsymbol{X}_j(t)^\top \boldsymbol{\beta}\right\}} \right]^{\mathrm{d}N_i(t)}, \tag{2}$$

where $Y_i(t) = I(T_i \geqslant t)$ is the at-risk indicator of the ith subject, $N_i(t)$ is the number of events for subject i at time t, and $\mathrm{d}N_i(t) = I(T_i \in [t, t + \Delta), \delta_i = 1)$, with Δ sufficiently small such that $\sum_{i=1}^{n} \mathrm{d}N_i(t) \leqslant 1$ for any t. Taking the natural logarithm of (2) gives the log

partial likelihood in the form of a summation:

$$pl(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^{\infty} \left[Y_i(t) \exp\left\{ \boldsymbol{X}_i(t)^\top \boldsymbol{\beta} \right\} - \log \sum_{j=1}^{n} Y_j(t) \exp\left\{ \boldsymbol{X}_j(t)^\top \boldsymbol{\beta} \right\} \right] \mathrm{d}N_i(t). \quad (3)$$

We differentiate $pl(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ to obtain the $p \times 1$ score vector, $\boldsymbol{U}(\boldsymbol{\beta})$:

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^{\infty} \left\{ \boldsymbol{X}_i(t) - \overline{\boldsymbol{X}}(\boldsymbol{\beta}, t) \right\} \mathrm{d}N_i(t),$$

where $\overline{\boldsymbol{X}}(\boldsymbol{\beta}, t)$ is a weighted mean of $\boldsymbol{X}_i$'s for those observations still at risk at time t with the weights being their corresponding risk scores, $\exp\{\boldsymbol{X}_i(t)^\top \boldsymbol{\beta}\}$. Taking the negative second order derivative of $pl(\boldsymbol{\beta})$ yields the observed information matrix $\boldsymbol{\mathcal{I}}_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^{\infty} \boldsymbol{V}(\boldsymbol{\beta}, t) \mathrm{d}N_i(t)$, with $\boldsymbol{V}(\boldsymbol{\beta}, t)$ being the weighted variance of $\boldsymbol{X}$ at time t:

$$\boldsymbol{V}(\boldsymbol{\beta}, t) = \frac{\sum_{i=1}^{n} Y_i(t) \exp\{\boldsymbol{X}_i(t)^\top \boldsymbol{\beta}\} \{\boldsymbol{X}_i(t) - \overline{\boldsymbol{X}}(\boldsymbol{\beta}, t)\} \{\boldsymbol{X}_i(t) - \overline{\boldsymbol{X}}(\boldsymbol{\beta}, t)\}^\top}{\sum_i Y_i(t) \exp\{\boldsymbol{X}_i(t)^\top \boldsymbol{\beta}\}}.$$

The maximum partial likelihood estimator $\widehat{\boldsymbol{\beta}}_n$ is obtained as the solution of $\boldsymbol{U}(\boldsymbol{\beta}) = \mathbf{0}$. The solution $\widehat{\boldsymbol{\beta}}_n$ is consistent, and asymptotically normal. The inverse of the observed information, $\boldsymbol{\mathcal{I}}_n(\widehat{\boldsymbol{\beta}}_n)$, is often used to approximate the asymptotic variance of $\widehat{\boldsymbol{\beta}}_n$.

2.2 *Test Statistic for Entire Dataset*

Following Grambsch and Therneau (1994), an alternative to PH in Model (1) is to allow time-varying coefficients, which can be characterized by

$$\beta_j(t) \equiv \beta_j + \theta_j g_j(t), \quad j = 1, \ldots, p, \quad (4)$$

where $g_j(t)$ is a function of time that varies around 0 and θ_j is a scalar. Common choices of $g(t)$ include the Kaplan–Meier (KM) transformation, which scales the horizontal axis by the left-continuous version of the KM survival curve, the identity function, and the natural logarithm function. Formulation (4) is rather general, as many tests fall within this framework for different choices of $g(t)$ (see, e.g., Xue and Schifano, 2017). Writing (4) in matrix notation yields

$$\lambda_i(t) = \lambda_0(t) \exp\left[\boldsymbol{X}_i(t)^\top \{\boldsymbol{\beta} + \boldsymbol{G}(t)\boldsymbol{\theta}\} \right], \quad i = 1, \ldots, n, \quad (5)$$

where $\boldsymbol{G}(t)$ is a $p \times p$ diagonal matrix with the jth diagonal element being $g_j(t)$, and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top$. Then the null hypothesis of $\boldsymbol{\beta}$ being time-invariant becomes $H_0 : \boldsymbol{\theta} = \mathbf{0}_{p\times 1}$.

The test of Grambsch and Therneau (1994) is based on Schoenfeld residuals. Assuming no tied event times and denoting them in increasing order as $t_1, \ldots, t_d$, where d is the total number of events among the n observations, the Schoenfeld residuals are defined as

$$\boldsymbol{r}_\ell(\boldsymbol{\beta}) = \boldsymbol{X}_{(\ell)} - \overline{\boldsymbol{X}}(\boldsymbol{\beta}, t_\ell),$$

where $\boldsymbol{X}_{(\ell)}$ is the covariate vector corresponding to the ℓth event time. In practice, we use $\widehat{\boldsymbol{\beta}}_n$ and obtain $\widehat{\boldsymbol{r}}_\ell$ for $\ell = 1, \ldots, d$. Let $\widehat{\boldsymbol{V}}_\ell = \boldsymbol{V}(\widehat{\boldsymbol{\beta}}_n, t_\ell)$, $\boldsymbol{G}_\ell = \boldsymbol{G}(t_\ell)$, and $\boldsymbol{H} = \sum_{\ell=1}^d \boldsymbol{G}_\ell \widehat{\boldsymbol{V}}_\ell \boldsymbol{G}_\ell - \left(\sum_{\ell=1}^d \boldsymbol{G}_\ell \widehat{\boldsymbol{V}}_\ell\right) \left(\sum_{\ell=1}^d \widehat{\boldsymbol{V}}_\ell\right)^{-1} \left(\sum_{\ell=1}^d \boldsymbol{G}_\ell \widehat{\boldsymbol{V}}_\ell\right)^\top$. Grambsch and Therneau (1994) proposed the statistic

$$T(\boldsymbol{G}) = \left(\sum_{\ell=1}^d \boldsymbol{G}_\ell \widehat{\boldsymbol{r}}_\ell\right)^\top \boldsymbol{H}^{-1} \left(\sum_{\ell=1}^d \boldsymbol{G}_\ell \widehat{\boldsymbol{r}}_\ell\right), \tag{6}$$

which, under the null hypothesis, has asymptotic distribution χ^2_p.

For identifiability, $g(t)$ is assumed to vary around 0, so for data analysis $\boldsymbol{G}_\ell$, $\ell = 1, \ldots, d$, need to be centered such that $\sum_{\ell=1}^d \boldsymbol{G}_\ell = \mathbf{0}_{p\times p}$. As pointed out by Therneau and Grambsch (2000), $\widehat{\boldsymbol{V}}_\ell$ is rather stable for most datasets, and therefore $\sum_{\ell=1}^d \boldsymbol{G}_\ell \widehat{\boldsymbol{V}}_\ell$ is often small. Therefore, $\boldsymbol{H}$ is often replaced by $\sum_{\ell=1}^d \boldsymbol{G}_\ell \widehat{\boldsymbol{V}}_\ell \boldsymbol{G}_\ell$. The `cox.zph()` function in the **survival** package implements the test in (6) using this same centering technique. In the sequel, we will assume that all $\boldsymbol{G}$ matrices are centered prior to any calculation of the diagnostic statistics.

Tied events are common in practice and there are several methods to handle ties. We use the approximation of Efron (1977), which is the default option in the package **survival** and returns fairly accurate results (Therneau and Grambsch, 2000, Section 3.3).

3. Online Updated Test and its Variations

3.1 *Cumulative Version*

Instead of a given, complete dataset, we now consider a scenario in which survival data become available in blocks. Suppose that for each new arriving block k, we observed d_k events among n_k subjects, for $k = 1, \ldots, K$, where K is some terminal accumulation point of interest. With a given $g(t)$ we obtain d_k centered $p \times p$ diagonal matrices $\boldsymbol{G}(t_1), \ldots, \boldsymbol{G}(t_{d_k})$ such that $\sum_{\ell=1}^{d_k} \boldsymbol{G}(t_\ell) = \boldsymbol{0}_{p\times p}$. Let $\boldsymbol{G}_{\ell k}$ and $\widehat{\boldsymbol{r}}_{\ell k}$, $\ell = 1, \ldots, d_k$, be the kth block counterpart of previously defined $\boldsymbol{G}_\ell$ and Schoenfeld residual $\widehat{\boldsymbol{r}}_\ell$, respectively. Without loss of generality, we assume that there is at least one event in each block so that a Cox model can be fitted, and each block-wise observed information matrix $\boldsymbol{\mathcal{I}}_{n_k,k}$, evaluated at some estimate of $\boldsymbol{\beta}$, is invertible. Let $\boldsymbol{V}_{\ell k}$ be the weighted variance-covariance matrix of the covariate matrix at the ℓth event time in the kth block. With the approximation that $\widehat{\boldsymbol{V}}_{\ell k} = \boldsymbol{\mathcal{I}}_{n_k,k}/d_k$, again where $\boldsymbol{\mathcal{I}}_{n_k,k}$ is evaluated at some estimate of $\boldsymbol{\beta}$, we have $\sum_{\ell=1}^{d_k} \boldsymbol{G}_{\ell k}\widehat{\boldsymbol{V}}_{\ell k} = \boldsymbol{0}_{p\times p}$. We will discuss the choice of estimate for $\boldsymbol{\beta}$ that will be used to evaluate $\boldsymbol{\mathcal{I}}_{n_k,k}$, and also $\widehat{\boldsymbol{r}}_{\ell k}$, in Section 3.3.

We denote $\boldsymbol{H}_{d_k,k} = (\sum_{\ell=1}^{d_k} \boldsymbol{G}_{\ell k}\boldsymbol{\mathcal{I}}_{n_k,k}\boldsymbol{G}_{\ell k})/d_k$, and $\boldsymbol{Q}_{d_k,k} = \sum_{\ell=1}^{d_k} \boldsymbol{G}_{\ell k}\widehat{\boldsymbol{r}}_{\ell k}$. Let $\boldsymbol{H}_0 = \boldsymbol{0}_{p\times p}$, $\boldsymbol{H}_{k-1} = \sum_{i=1}^{k-1} \boldsymbol{H}_{d_i,i}$, $\boldsymbol{Q}_0 = \boldsymbol{0}_{p\times 1}$, and $\boldsymbol{Q}_{k-1} = \sum_{i=1}^{k-1} \boldsymbol{Q}_{d_i,i}$. Then we have the online updating test statistic given by

$$T_k(\boldsymbol{G}) = \boldsymbol{Q}_k^\top \boldsymbol{H}_k^{-1}\boldsymbol{Q}_k = (\boldsymbol{Q}_{k-1} + \boldsymbol{Q}_{d_k,k})^\top (\boldsymbol{H}_{k-1} + \boldsymbol{H}_{d_k,k})^{-1}(\boldsymbol{Q}_{k-1} + \boldsymbol{Q}_{d_k,k}). \tag{7}$$

At each accumulation point k, we need to store $\boldsymbol{H}_{k-1}$ and $\boldsymbol{Q}_{k-1}$ from previous calculations, and compute $\boldsymbol{H}_{d_k,k}$ and $\boldsymbol{Q}_{d_k,k}$ for the current block.

3.2 *Window Version*

The cumulative test statistic takes all historical blocks into consideration, one potential problem of which is that discrepancies from the PH assumption will accumulate and after a certain time period, the test will always reject the null hypothesis. This motivates us to focus on more recent blocks in some applications. At block k, we consider a window of width $w(\geqslant$

1), which is tunable, and use summary statistics for all blocks in this window to construct the corresponding test statistic. With $\boldsymbol{H}_{d_k,k}$ and $\boldsymbol{Q}_{d_k,k}$ defined above, we again assume there is at least one event in each block of data. Denoting $\boldsymbol{H}_k^w = \sum_{i=k+1-w}^{k} \boldsymbol{H}_{d_i,i}$, and $\boldsymbol{Q}_k^w = \sum_{i=k+1-w}^{k} \boldsymbol{Q}_{d_i,i}$, the window version online updating test statistic for nonproportionality based on the most recent w blocks is:

$$T_k^w(\boldsymbol{G}) = (\boldsymbol{Q}_k^w)^\top (\boldsymbol{H}_k^w)^{-1} \boldsymbol{Q}_k^w. \tag{8}$$

In implementation, we only need to store $\boldsymbol{H}_{d_k,k}$ and $\boldsymbol{Q}_{d_k,k}$ for all but the first block in the window, and compute these summary statistics for the current block to obtain the aggregated diagnostic statistic. Compared to the cumulative version statistic, which at each update requires storage of one $p \times 1$ vector $\boldsymbol{Q}_k$, one $p \times 1$ vector for an estimate of $\boldsymbol{\beta}$, one $p \times p$ matrix $\boldsymbol{H}_k$, and one $p \times p$ variance matrix of $\boldsymbol{\beta}$, the window version requires storage of these quantities for $w - 1$ steps, which is still minimally storage intensive when $p \ll n_k$. In addition, as an auxiliary approach that provides an indication approximately where along the stream a violation has occurred, w is generally chosen to not be large. This also makes the storage of these quantities affordable, and the handling of large blocks possible.

3.3 *Where to Evaluate the Matrices and Residuals*

The observed information matrix $\boldsymbol{\mathcal{I}}_{n_k,k}$ and the residuals $\widehat{\boldsymbol{r}}_{\ell k}$ must be evaluated at a particular choice of $\boldsymbol{\beta}$. A straightforward choice would be $\widehat{\boldsymbol{\beta}}_{n_k,k}$, the estimate of $\boldsymbol{\beta}$ using the kth block of data, $k = 1, 2, \ldots, K$. It may, however, be more advantageous to use an estimate that utilizes all relevant historical information.

Now let us consider the kth accumulation point. The score function for subset k can be obtained as $\boldsymbol{U}_{n_k,k}(\boldsymbol{\beta})$, and we denote the solution to $\boldsymbol{U}_{n_k,k}(\boldsymbol{\beta}) = \boldsymbol{0}_{p\times 1}$ as $\widehat{\boldsymbol{\beta}}_{n_k,k}$. A Taylor expansion of $-\boldsymbol{U}_{n_k,k}(\boldsymbol{\beta})$ at $\widehat{\boldsymbol{\beta}}_{n_k,k}$ is given by

$$-\boldsymbol{U}_{n_k,k}(\boldsymbol{\beta}) = \boldsymbol{\mathcal{I}}_{n_k,k}(\widehat{\boldsymbol{\beta}}_{n_k,k})(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{n_k,k}) + \boldsymbol{R}_{n_k,k}$$

as $\boldsymbol{U}_{n_k,k}(\widehat{\boldsymbol{\beta}}_{n_k,k}) = \mathbf{0}_{p\times 1}$ and $\boldsymbol{R}_{n_k,k}$ is the remainder term. Again, without loss of generality, we assume that there is at least one event in each block, and each $\boldsymbol{\mathcal{I}}_{n_k,k}$ is invertible.

Denote $\boldsymbol{\mathcal{I}}_{n_k,k}(\widehat{\boldsymbol{\beta}}_{n_k,k})$ as $\widehat{\boldsymbol{\mathcal{I}}}_{n_k,k}$. Similar to the aggregated estimating equation (AEE) estimator of Lin and Xi (2011), which uses a weighted combination of the subset estimators, an AEE estimator under the Cox model framework may be given by

$$\widehat{\boldsymbol{\beta}}_N = \left(\sum_{k=1}^{K} \widehat{\boldsymbol{\mathcal{I}}}_{n_k,k}\right)^{-1} \sum_{k=1}^{K} \widehat{\boldsymbol{\mathcal{I}}}_{n_k,k}\widehat{\boldsymbol{\beta}}_{n_k,k}, \tag{9}$$

which is the solution to $\sum_{k=1}^{K} \widehat{\boldsymbol{\mathcal{I}}}_{n_k,k}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{n_k,k}) = \mathbf{0}_{p\times 1}$, with N being the total number of observations at the final accumulation point K. Schifano et al. (2016) provided the variance estimator for the original AEE estimator of Lin and Xi (2011), and under the Cox model framework it simplifies to $\widehat{\boldsymbol{A}}_N = \left(\sum_{k=1}^{K} \widehat{\boldsymbol{\mathcal{I}}}_{n_k,k}\right)^{-1}$.

Following Schifano et al. (2016), a cumulative estimating equation (CEE) estimator for $\boldsymbol{\beta}$ at accumulation point k under the Cox model framework is

$$\widehat{\boldsymbol{\beta}}_k = \left(\widehat{\boldsymbol{\mathcal{I}}}_{k-1} + \widehat{\boldsymbol{\mathcal{I}}}_{n_k,k}\right)^{-1} \left(\widehat{\boldsymbol{\mathcal{I}}}_{k-1}\widehat{\boldsymbol{\beta}}_{k-1} + \widehat{\boldsymbol{\mathcal{I}}}_{n_k,k}\widehat{\boldsymbol{\beta}}_{n_k,k}\right) \tag{10}$$

for $k = 1, 2, \ldots$, where $\widehat{\boldsymbol{\beta}}_0 = \mathbf{0}_{p\times 1}$, $\widehat{\boldsymbol{\mathcal{I}}}_0 = \mathbf{0}_{p\times p}$, and $\widehat{\boldsymbol{\mathcal{I}}}_k = \sum_{i=1}^{k} \widehat{\boldsymbol{\mathcal{I}}}_{n_i,i} = \widehat{\boldsymbol{\mathcal{I}}}_{k-1} + \widehat{\boldsymbol{\mathcal{I}}}_{n_k,k}$. The variance estimator at the kth update simplifies to $\widehat{\boldsymbol{A}}_k = \left(\widehat{\boldsymbol{\mathcal{I}}}_{k-1} + \widehat{\boldsymbol{\mathcal{I}}}_{n_k,k}\right)^{-1}$. Note that for terminal $k = K$, the AEE estimators and CEE estimators coincide.

Similar to Schifano et al. (2016), we propose a CUEE estimator framework to better approximate the maximum partial likelihood estimator (based on the entire sample) with less bias. Take the Taylor expansion of $-\boldsymbol{U}_{n_k,k}(\boldsymbol{\beta})$ around $\breve{\boldsymbol{\beta}}_{n_k,k}$, which will be defined later. We have

$$-\boldsymbol{U}_{n_k,k}(\boldsymbol{\beta}) = -\boldsymbol{U}_{n_k,k}(\breve{\boldsymbol{\beta}}_{n_k,k}) + \boldsymbol{\mathcal{I}}_{n_k,k}(\breve{\boldsymbol{\beta}}_{n_k,k})(\boldsymbol{\beta} - \breve{\boldsymbol{\beta}}_{n_k,k}) + \breve{\boldsymbol{R}}_{n_k,k},$$

where $\breve{\boldsymbol{R}}_{n_k,k}$ is the remainder term. Again for simplicity, we denote $\boldsymbol{\mathcal{I}}_{n_k,k}(\breve{\boldsymbol{\beta}}_{n_k,k})$ as $\breve{\boldsymbol{\mathcal{I}}}_{n_k,k}$, and $\boldsymbol{U}(\breve{\boldsymbol{\beta}}_{n_k,k})$ as $\breve{\boldsymbol{U}}_{n_k,k}$. We now ignore the remainder term and sum the first order expansions

for blocks $1, \ldots, K$, and set it equal to $\mathbf{0}_{p\times 1}$:

$$\sum_{k=1}^{K} -\breve{\boldsymbol{U}}_{n_k,k} + \sum_{k=1}^{K} \breve{\boldsymbol{\mathcal{I}}}_{n_k,k} \left(\boldsymbol{\beta} - \breve{\boldsymbol{\beta}}_{n_k,k}\right) = \mathbf{0}_{p\times 1}. \tag{11}$$

Then we have the solution to (11): $\widetilde{\boldsymbol{\beta}}_K = \left(\sum_{k=1}^{K} \breve{\boldsymbol{\mathcal{I}}}_{n_k,k}\right)^{-1} \left(\sum_{k=1}^{K} \breve{\boldsymbol{\mathcal{I}}}_{n_k,k}\breve{\boldsymbol{\beta}}_{n_k,k} + \sum_{k=1}^{K} \breve{\boldsymbol{U}}_{n_k,k}\right)$. The choice of $\breve{\boldsymbol{\beta}}_{n_k,k}$ is subjective. At accumulation point k, it is possible to utilize information at the previous accumulation point to define $\breve{\boldsymbol{\beta}}_{n_k,k}$. One candidate intermediary estimator is

$$\breve{\boldsymbol{\beta}}_{n_k,k} = \left(\breve{\boldsymbol{\mathcal{I}}}_{k-1} + \widehat{\boldsymbol{\mathcal{I}}}_{n_k,k}\right)^{-1} \left(\sum_{i=1}^{k-1} \breve{\boldsymbol{\mathcal{I}}}_{n_i,i}\breve{\boldsymbol{\beta}}_{n_i,i} + \widehat{\boldsymbol{\mathcal{I}}}_{n_k,k}\widehat{\boldsymbol{\beta}}_{n_k,k}\right) \tag{12}$$

for $k = 1, 2, \ldots$, $\breve{\boldsymbol{\mathcal{I}}}_0 = \mathbf{0}_{p\times p}$, $\breve{\boldsymbol{\beta}}_{n_0,0} = \mathbf{0}_{p\times 1}$, and $\breve{\boldsymbol{\mathcal{I}}}_k = \sum_{i=1}^{k} \breve{\boldsymbol{\mathcal{I}}}_{n_i,i}$. Estimator (12) is the weighted combination of the previous intermediary estimators $\breve{\boldsymbol{\beta}}_{n_i,i}, i = 1, \ldots, k-1$ and the current subset estimator $\widehat{\boldsymbol{\beta}}_{n_k,k}$. It results as the solution to the estimating equation $\sum_{i=1}^{k-1} \breve{\boldsymbol{\mathcal{I}}}_{n_i,i} \left(\boldsymbol{\beta} - \breve{\boldsymbol{\beta}}_{n_i,i}\right) + \widehat{\boldsymbol{\mathcal{I}}}_{n_k,k} \left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{n_k,k}\right) = \mathbf{0}_{p\times 1}$, with $\widehat{\boldsymbol{\mathcal{I}}}_{n_k,k} \left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{n_k,k}\right)$ being the bias correction term since $-\sum_{i=1}^{k-1} \breve{\boldsymbol{U}}_{n_i,i}$ has been omitted.

With $\breve{\boldsymbol{\beta}}_{n_k,k}$ given in (12), the CUEE estimator $\widetilde{\boldsymbol{\beta}}_k$ for the Cox model is

$$\widetilde{\boldsymbol{\beta}}_k = \left(\breve{\boldsymbol{\mathcal{I}}}_{k-1} + \breve{\boldsymbol{\mathcal{I}}}_{n_k,k}\right)^{-1} \left(\boldsymbol{s}_{k-1} + \breve{\boldsymbol{\mathcal{I}}}_{n_k,k}\breve{\boldsymbol{\beta}}_{n_k,k} + \boldsymbol{\xi}_{k-1} + \breve{\boldsymbol{U}}_{n_k,k}\right),$$

with $\boldsymbol{s}_k = \sum_{i=1}^{k} \breve{\boldsymbol{\mathcal{I}}}_{n_i,i}\breve{\boldsymbol{\beta}}_{n_i,i} = \breve{\boldsymbol{\mathcal{I}}}_{n_k,k}\breve{\boldsymbol{\beta}}_{n_k,k} + s_{k-1}$ and $\boldsymbol{\xi}_k = \sum_{i=1}^{k} \breve{\boldsymbol{U}}_{n_i,i} = \breve{\boldsymbol{U}}_{n_k,k} + \boldsymbol{\xi}_{k-1}$, where $\boldsymbol{s}_0 = \boldsymbol{\xi}_0 = \mathbf{0}_{p\times 1}$, and $k = 1, 2, \ldots$. For the variance of $\widetilde{\boldsymbol{\beta}}_k$, as $\mathbf{0}_{p\times 1} = -\widehat{\boldsymbol{U}}_{n_k,k} \approx -\breve{\boldsymbol{U}}_{n_k,k} + \widehat{\boldsymbol{\mathcal{I}}}_{n_k,k} \left(\widehat{\boldsymbol{\beta}}_{n_k,k} - \breve{\boldsymbol{\beta}}_{n_k,k}\right)$, we have $\breve{\boldsymbol{\mathcal{I}}}_{n_k,k}\breve{\boldsymbol{\beta}}_{n_k,k} + \breve{\boldsymbol{U}}_{n_k,k} \approx \breve{\boldsymbol{\mathcal{I}}}_{n_k,k}\widehat{\boldsymbol{\beta}}_{n_k,k}$. The estimated variance of $\widetilde{\boldsymbol{\beta}}_k$ is online updated by, in simplified form,

$$\widetilde{\mathrm{Var}}(\widetilde{\boldsymbol{\beta}}_k) = \left(\breve{\boldsymbol{\mathcal{I}}}_{k-1} + \breve{\boldsymbol{\mathcal{I}}}_{n_k,k}\right)^{-1} \left(\sum_{i=1}^{k} \breve{\boldsymbol{\mathcal{I}}}_{n_k,k}\widehat{\boldsymbol{\mathcal{I}}}_{n_k,k}^{-1}\breve{\boldsymbol{\mathcal{I}}}_{n_k,k}^{\top}\right) \left\{\left(\breve{\boldsymbol{\mathcal{I}}}_{k-1} + \breve{\boldsymbol{\mathcal{I}}}_{n_k,k}\right)^{-1}\right\}^{\top}.$$

Thus, for the cumulative version statistic, the matrices and Schoenfeld residuals are evaluated at $\widetilde{\boldsymbol{\beta}}_k$, the CUEE estimator, in our implementation. For the window version statistic, the matrices and Schoenfeld residuals are evaluated at the CEE estimator, as with a limited window size, there is little room for the bias of the CEE estimator to accumulate, and the difference between the CUEE estimator and the CEE estimator within a window is

negligible for small w. Note that when $w = 1$, both estimators are the same, and are equal to the parameter estimate for the current block, $\widehat{\boldsymbol{\beta}}_{n_k,k}$.

3.4 *Asymptotic Results*

We now provide the asymptotic distribution of the test statistic $T_k(\boldsymbol{G})$ given in Equation (7). For ease of presentation, we assume that all subsets of data are of equal size n, i.e., $n_k = n$.

THEOREM 3.1: *Under conditions C1-C5 in Web Appendix A, as* $n \to \infty$, *if* $K = O(n^\gamma)$ *with* $0 < \gamma < \min\{1 - 2\alpha, 4\alpha - 1\}$, *then for any* $k \leqslant K$, *the test statistic satisfies that*

$$T_k(\boldsymbol{G}) \to \chi^2_p,$$

in distribution when all data blocks follow the PH model with the same covariate parameters.

The proof is provided in Web Appendix A. The asymptotic distribution is valid for any stage of the updating process if each subset is not very small and the null hypothesis is true. This means that the type one error rate is always well maintained. As more data accumulate along the updating procedure, the test statistic gains more power. If the n_k's are different, the asymptotic result is still valid under some mild condition, for example, $\max_k n_k / \min_k n_k = O(1)$. Note that the window version statistic $T^w_k(\boldsymbol{G})$ is essentially the cumulative version statistic evaluated at the CEE with different starting blocks. Therefore, the asymptotic distribution is also valid for the window version statistic. In the special case of $w = 1$, the proposed statistic reduces to the original $T(\boldsymbol{G})$ on the most recent block, which has been shown to be χ^2_p by Grambsch and Therneau (1994).

4. Simulation Studies

Simulation studies were carried out to evaluate the empirical sizes and powers of both $T_k(\boldsymbol{G})$ and $T^w_k(\boldsymbol{G})$. When data were generated under the PH assumption, we also compared the empirical distribution of $T_k(\boldsymbol{G})$ with that of the standard statistic computed using all data

up to selective accumulation points k, denoted by $T_{1:k}(\boldsymbol{G})$. While we look at the end of each stream to decide whether the entire stream of data satisfies the PH assumption or not, we also examine the results at each accumulation point to verify the performance of the proposed test statistics. Simulations have also been conducted to assess the savings in computing time and reduction in memory usage for the proposed statistics with big survival data. See Web Appendices B.1 and B.2.

4.1 *Size*

Event times were generated from Model (1) with three covariates $x_{ki[1]} \overset{\text{i.i.d.}}{\sim} N(0,1)$, $x_{ki[2]} \overset{\text{i.i.d.}}{\sim}$ Bernoulli(0.5), $x_{ki[3]} \overset{\text{i.i.d.}}{\sim}$ Bernoulli(0.1) for $i = 1, \ldots, n_k$, making a $n_k \times 3$ covariate matrix. We set a vector of parameters $\boldsymbol{\beta}_0 = (0.67, -0.26, 0.36)^\top$, and baseline hazard $\lambda_0(t) = 0.018$. Censoring times were generated independently from a mixture distribution: $\varepsilon\langle 60\rangle + (1-\varepsilon)\mathscr{U}(0,60)$, where $\langle 60\rangle$ represents a point mass at 60, and $\mathscr{U}(0,60)$ denotes the uniform distribution over (0, 60). Setting $\varepsilon = 0.9$ gives approximately 40% censoring rate, and $\varepsilon = 0.1$ gives approximately 60% censoring rate. For each censoring level, we generated $1,000$ independent streams of survival datasets, each of which had $N = 200,000$ observations in $K = 100$ blocks with $n_k = 2,000$.

[Figure 1 about here.]

Three choices of $g(t)$ were considered, the identity, KM, and log transformations, in the calculation of the test statistics. For each choice, we calculated both $T_k(\boldsymbol{G})$ and $T_k^w(\boldsymbol{G})$ with $w = 5$ upon arrival of each block of simulated data. Figure 1 summarizes empirical sizes of the test with nominal level 0.05 at each accumulation point $k = 1, \ldots, 100$ for the two versions of the tests under two censoring levels. The empirical sizes for the three choices of $g(t)$ fluctuate closely around the nominal level 0.05 in all the scenarios. The log transformation, however, results in a slightly larger size, and its usage should therefore be treated with caution.

[Figure 2 about here.]

To compare the empirical distribution of $T_k(\boldsymbol{G})$ and the standard statistic $T_{1:k}(\boldsymbol{G})$, we additionally computed $T_{1:k}(\boldsymbol{G})$ at blocks $k \in \{25, 50, 75, 100\}$ based on cumulative data up to those blocks. Figure 2 presents the quantile-quantile plots of the two statistics obtained with $g(t)$ being the KM transformation. The points line up closely on the 45 degree line, confirming that the online updating cumulative statistics $T_k(\boldsymbol{G})$ follow the same asymptotic χ^2_p distribution under the null hypothesis as $T_{1:k}(\boldsymbol{G})$.

Additional simulation results on the sizes for scenarios where $p \in \{10, 20\}$ and where covariate coefficients are piecewise constant with respect to time (and accommodated in the PH model by including additional covariates to handle the pieces separately) are reported in Web Appendices B.3 and B.4. In both cases, the size was well-maintained.

4.2 *Power*

Continuing with the simulation setting from Section 4.1, two scenarios where the PH assumption is violated were considered to assess the power of the proposed tests.

[Figure 3 about here.]

The first scenario breaks the PH assumption by a multiplicative frailty in the hazard function. Starting from the 51st block in each stream, the hazard function, instead of being (1), becomes $\lambda_i(t) = \lambda_0(t) \exp\big(X_i^\top \boldsymbol{\beta} + \epsilon_i\big)$, where a normal frailty $\epsilon_i \sim N(0, \sigma^2)$ is introduced. Two levels of σ were considered, 0.5 and 1. Figure 3 shows the empirical rejection rates of the tests at level 0.05 from 1,000 replicates against accumulation point k. The tests have higher power under lower censoring rate or higher frailty standard deviation. At a given censoring rate and frailty standard deviation, $T_k^w(\boldsymbol{G})$ picks up the change more rapidly than $T_k(\boldsymbol{G})$ because it discards information from older blocks for which the PH assumption holds; the power remains at a certain level (less than 1) after all the blocks in the window

contain data generated from the frailty model. While $T_k(\boldsymbol{G})$ responds to the change more slowly, as the proportion of blocks with data generated from the frailty model increases, the power approaches 1 eventually. In all settings, tests based on the log and KM transformations seem to have higher power than that based on the identity transformation.

The second scenario breaks the PH assumption by a change in one of the covariate effects. Specifically, we considered an increase of 0.5 or 1 in β_1, the coefficient for the first covariate in data generation, starting from the 51st block. The empirical rejection rates of the tests with level 0.05 from 1,000 replicates are presented in Figure 3. Both versions of the tests have higher power when the censoring rate is lower or the change in β_1 is larger. At a given censoring rate and change in β_1, $T_k^w(\boldsymbol{G})$ only has power to detect the change near the 51st block, where the blocks in the window contain data from two models. The cumulative version, $T_k(\boldsymbol{G})$, picks up the change after the 51st block and the power increases quickly to 1.

A more comprehensive simulation study was conducted to compare the power of $T_K(\boldsymbol{G})$ and the full data test statistic $T(\boldsymbol{G}) = T_{1:K}(\boldsymbol{G})$ at the end of each data stream, and the results are presented in Web Appendix B.5. When there is a model change, the power of $T_K(\boldsymbol{G})$ is comparable to the power of $T(\boldsymbol{G})$; when there is a change in covariate effect, $T_K(\boldsymbol{G})$ has significantly higher power than $T(\boldsymbol{G})$.

5. Survival Analysis of SEER Lymphoma Patients

We consider analyzing the survival time of the lymphoma patients in the SEER program with the proposed methods. Among the 131,960 patients diagnosed with lymphoma between 1973 to 2007, 47,009 experienced an event within 60 months due to lymphoma, resulting in a censoring rate of 64.4%. The risk factors considered in our analysis were Age (centered and scaled), gender indicator (Female), African-American indicator (Black). There were 60,432 females, and 9,199 African-Americans. We wish to compare the performance of the standard statistic $T(\boldsymbol{G})$ from Equation (6) with $T_k(\boldsymbol{G})$ under a setting in which the PH assumption

is judged to be satisfied based on the standard $T(\boldsymbol{G})$ test. For online updating, the patients in the data were ordered by time of diagnosis, and partitioned by quarter of a year into 140 blocks. The average sample size per block was 943, but the block sizes and censoring rates increased over time; see Web Figure 4.

As a starting point, an initial model that included the three risk factors was fitted, and $T(\boldsymbol{G})$ based on the full data as in Equation (6) was calculated to be 83.38, which indicated that the model does not satisfy the PH assumption. The online updating cumulative statistic $T_k(\boldsymbol{G})$ was calculated to be 95.60. Due to the relatively high censoring rate, the KM transformation was chosen in calculation of the diagnostic statistics as it is more robust in such a scenario (e.g., Xue and Schifano, 2017). Diagnosis with function plot.cox.zph() in the **survival** package revealed that all the parameters are likely to be time-dependent; see Web Figure 5.

Techniques in Therneau et al. (2018) were used to allow the parameters to be piecewise-constant over time. Two cut-offs were chosen at 2 and 30 months based on the time-variation pattern of $\widehat{\boldsymbol{\beta}}(t)$ obtained from the naive model. A factor variable tgroup is defined to indicate on which intervals the corresponding observation contributes to estimation of $\boldsymbol{\beta}$. For example, a subject with survival time 25 and event 1 will now be represented separately on two intervals: one with time interval (0,2], with event 0 and tgroup = 1, and the other with time interval (2,25], with event 1 and tgroup = 2. The interaction of Age, Female and Black with the generated tgroup as strata gives the model more flexibility to fit to the data. The new model resulted in $T(\boldsymbol{G}) = T_{1:140}(\boldsymbol{G}) = 5.75$ on 9 degrees of freedom with a p-value of 0.77, which indicates that the PH assumption for the revised model is appropriate based on the full data. Web Figure 6 presents time-variation plot of parameters for the revised model.

To evaluate the performance of the online updating parameter estimates and test statistics under the revised model, at each block k, $k = 1, \ldots, 140$, we calculated the parameter

estimates, $T_k(\boldsymbol{G})$, $T_k^w(\boldsymbol{G})$, and also $T_{1:k}(\boldsymbol{G})$ based on the single large dataset consisting of all cumulative data up to block k. Two versions of $T_k(\boldsymbol{G})$ were obtained, one using the CEE estimator $\widehat{\boldsymbol{\beta}}_k$ and the other using the CUEE estimator $\widetilde{\boldsymbol{\beta}}_k$. For $T_k^w(\boldsymbol{G})$, the CEE estimator $\widehat{\boldsymbol{\beta}}_k$ was used as discussed previously, and two widths $w = 1$ and $w = 10$ were considered. The trajectories of different versions of the test statistics were plotted in the left panel of Figure 4. While the PH assumption seemed to be satisfied within each individual block ($w = 1$), as well as in cumulative data up to each accumulation point, both online updating cumulative statistics $T_k(\boldsymbol{G})$ resulted in a rejection of the null hypothesis, and $T_k^w(\boldsymbol{G})$ when $w = 10$ also resulted in a few rejections along the stream.

[Figure 4 about here.]

[Figure 5 about here.]

The trajectories of three parameter estimates $\widehat{\beta}_{\mathsf{Age}}$, $\widehat{\beta}_{\mathsf{Female}}$, and $\widehat{\beta}_{\mathsf{Black}}$ on the three time intervals $(0, 2]$, $(2, 30]$ and $(30, 60]$ (obtained from the covariate interactions with tgroup) were plotted with respect to block indices to investigate this apparent discrepancy; see Figure 5. Apparently, $\widehat{\beta}_{\mathsf{Age}}$ on $(0, 2]$ remained relatively stable for blocks 1 to 50, but started to first decrease and later increase. This change was captured by both $T_k^w(\boldsymbol{G})$ and $T_k(\boldsymbol{G})$, but not by $T_{1:k}(\boldsymbol{G})$. This is explained by the fact that $T_{1:k}(\boldsymbol{G})$ is based on a single estimator of $\boldsymbol{\beta}$, while in the online updating statistics, each block has its own estimate of $\boldsymbol{\beta}$. The temporal changes that are observed in the CUEE estimate of $\boldsymbol{\beta}$ get canceled in the calculation based on the full cumulative data.

[Figure 6 about here.]

To confirm that the temporal change in parameter contributed to the highly significant online updating test statistics, we randomly permuted the order of the observations in the original dataset 1,000 times using the same block size as the original data. For each permuta-

tion, we applied the same techniques and cut-offs to allow for piecewise constant parameters over time as before. The histogram of the 1,000 CUEE-based $T_k(\boldsymbol{G})$ is included in Web Figure 7. The empirical p-value based on these 1,000 permutations is 0.016, indicating that the particular order of blocks in the original temporally ordered data is indeed contributing to non-proportionality. Figure 6 presents the same diagnostic plots as Figure 5 except that they are for one random permutation. While the final cumulative data parameter estimates remain the same, the trajectories are much flatter, with no obvious temporal trend over blocks. The diagnostic statistics were also obtained under this random permutation, and plotted in the right panel of Figure 4. Each block again satisfies the PH assumption, and the performance of the online updating cumulative statistic based on CUEE is very close to $T(\boldsymbol{G})$ computed on the entire dataset. The online updating window version ($w = 10$), however, still identified a few neighborhoods where the variation is large, and this behavior persists across different choices of window size.

6. Discussion

We developed online updating test statistics for the PH assumption of the Cox model for streams of survival data. The test statistics were inspired by the divide and conquer approach (Lin and Xi, 2011) and the online updating approach for estimation and inference of regression parameters for estimating equations (Schifano et al., 2016). We proposed two versions of test statistics, $T_k(\boldsymbol{G})$ using cumulative information from all historical data, and $T_k^w(\boldsymbol{G})$ using information only from more recent data. Both statistics have an asymptotic χ_p^2 distribution under the null hypothesis. In our simulation studies, the power of $T_k(\boldsymbol{G})$ is comparable to or higher than the power of the standard test $T(\boldsymbol{G})$ on the entire dataset, for scenarios of a model change or parameter change, respectively. In addition, when $T(\boldsymbol{G})$ fails to detect violation of the null hypothesis on the whole dataset, $T_k(\boldsymbol{G})$ may still identify the violation with high power. This was observed in our application to the SEER data, and also

echoes the findings in Battey et al. (2018). This also suggests that, even when the dataset is not huge, it might be desirable to partition the data and examine the partitions for possibly masked violations of the null hypothesis. At the final block, the cumulative version test statistic will help us decide if the PH assumption has been satisfied. The window version, however, can be run at the same time, as it is sensitive to heterogeneity among a few blocks.

As with previous online updating approaches, $T_k(\boldsymbol{G})$ and $T_k^w(\boldsymbol{G})$ are computationally fast, and minimally storage intensive. As shown in the supporting information, the methods are also capable of handling large datasets of a few gigabyte's size, and can return the estimation and diagnostic results within reasonable time limit. Compared to parallel computing for such datasets, the proposed approach reduces time needed for communication between nodes, and allows for bias correction of the parameter estimates.

A few issues beyond the scope of this paper are worth further investigation. The size of blocks should be chosen following general guidelines (e.g., Schoenfeld, 1983) so that the covariate effects can be sufficiently identified, and that the information matrices exist and are invertible. In practice, with a data stream, we can always choose to let the data accumulate until a certain number of events are observed. Then these observations can be grouped into one block, which can produce stable and valid results for test purposes. For $T_k^w(\boldsymbol{G})$, the choice of w may affect the test results and local parameter estimates. Possible influential factors include the size of data chunks, the censoring rate within each chunk, among others. Additionally, as we are more interested in local or current goodness-of-fit when using the window version, w should generally be small. Also, as illustrated in Figure 3, $T_k^w(\boldsymbol{G})$ can behave differently under different violations of the PH assumption, therefore, prior knowledge on what types of changes are likely to occur, if available, may also be taken into consideration. As we are more concerned with deciding whether the entire stream satisfies the PH assumption, this window version should be treated as of auxiliary purpose.

Also, the test statistics and parameter estimates perform well when p is small to moderate. When p is high or ultra-high, singularity issues could arise, and appropriate penalization methods should be considered (e.g. Fan and Li, 2002; Zou, 2008; Mittal et al., 2014).

Finally, in this work we are only concerned with making a final decision regarding the PH assumption at the end of a data stream. There are scenarios, however, under which we may wish to make decisions alongside the data stream as the updating process progresses. This brings up the issue of multiple hypothesis testing. Hypothesis testing in the online updating framework is an interesting topic, and has been explored recently in Webb and Petitjean (2016) and Javanmard and Montanari (2018), and also in the statistical process control framework in, e.g., Lee and Jun (2010, 2012). Appropriate adjustment procedures in the online updating PH test context are areas devoted for future research.

References

Air Transport Action Group (2018). Aviation: Benefits beyond borders (2018) – global summary. https://www.atag.org/component/attachments/attachments.html?id=708. Online; accessed Dec 30, 2018.

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

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

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

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association* **72,** 557–565.

Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Annals of Statistics* **30,** 74–99.

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

Grambsch, P. M. and Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* **81,** 515–526.

Javanmard, A. and Montanari, A. (2018). Online rules for control of false discovery rate and false discovery exceedance. *The Annals of Statistics* **46,** 526–554.

Lee, S.-H. and Jun, C.-H. (2010). A new control scheme always better than X-bar chart. *Communications in Statistics – Theory and Methods* **39,** 3492–3503.

Lee, S.-H. and Jun, C.-H. (2012). A process monitoring scheme controlling false discovery rate. *Communications in Statistics – Simulation and Computation* **41,** 1912–1920.

Lin, N. and Xi, R. (2011). Aggregated estimating equation estimation. *Statistics and its Interface* **4,** 73–83.

Mittal, S., Madigan, D., Burd, R. S., and Suchard, M. A. (2014). High-dimensional, massive sample-size Cox proportional hazards regression for survival analysis. *Biostatistics* **15,** 207–221.

National Association of Realtors (2018). Quick real estate statistics. https://www.nar.realtor/research-and-statistics/quick-real-estate-statistics. Online; accessed Dec 30, 2018.

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

Schoenfeld, D. A. (1983). Sample-size formula for the proportional-hazards regression model. *Biometrics* **39,** 499–503.

Therneau, T., Crowson, C., and Atkinson, E. (2018). Using time dependent covariates and time dependent coefficients in the Cox model.

Therneau, T. M. (2015). *A Package for Survival Analysis in S*. version 2.38.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag Inc, Berlin; New York.

Wang, Y., Palmer, N., Di, Q., Schwartz, J., Kohane, I., and Cai, T. (2019). A fast divide-and-conquer sparse Cox regression. *Biostatistics* Forthcoming.

Webb, G. I. and Petitjean, F. (2016). A multiple test correction for streams and cascades of statistical hypothesis tests. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1255–1264, New York, NY, USA. ACM.

Xue, Y. (2018). ys-xue/Code-for-Online-Updating-Proportional -Hazards-Test:First Release.

Xue, Y. and Schifano, E. D. (2017). Diagnostics for the Cox model. *Communications for Statistical Applications and Methods* **24,** 583–604.

Zou, H. (2008). A note on path-based variable selection in the penalized proportional hazards model. *Biometrika* **95,** 241–247.

Supporting Information

Web Appendices, Tables, and Figures referenced in Sections 3 to 5 are available with this paper at the Biometrics website on Wiley Online Library. Method implementation in R, as well as an example using a simulated dataset, can be found at https://github.com/ys-xue/Code-for-Online-Updating-Proportional-Hazards-Test.

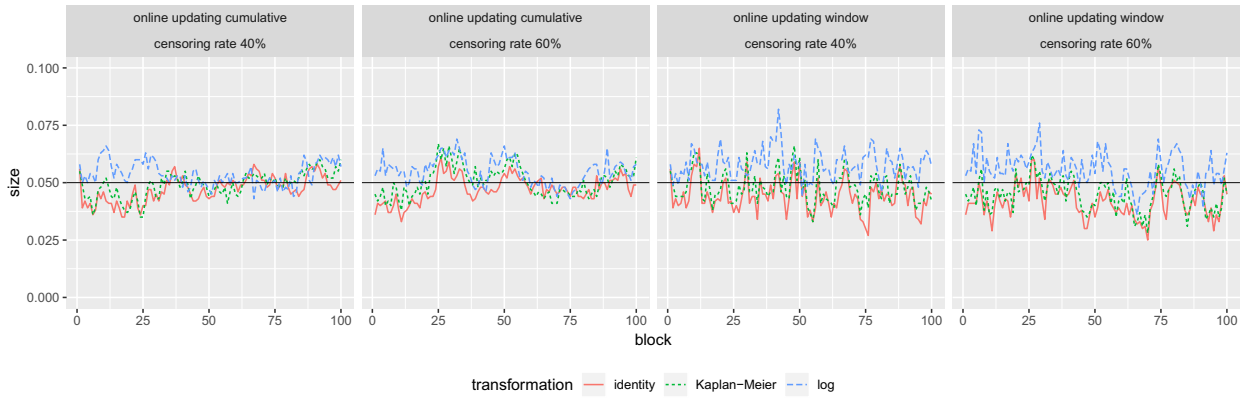


Figure 1. Empirical size (proportion of statistic values greater than $\chi^2_{3,0.95}$) calculated at each update using the identity, KM, and log transformations under the null hypothesis. This figure appears in color in the electronic version of this article.

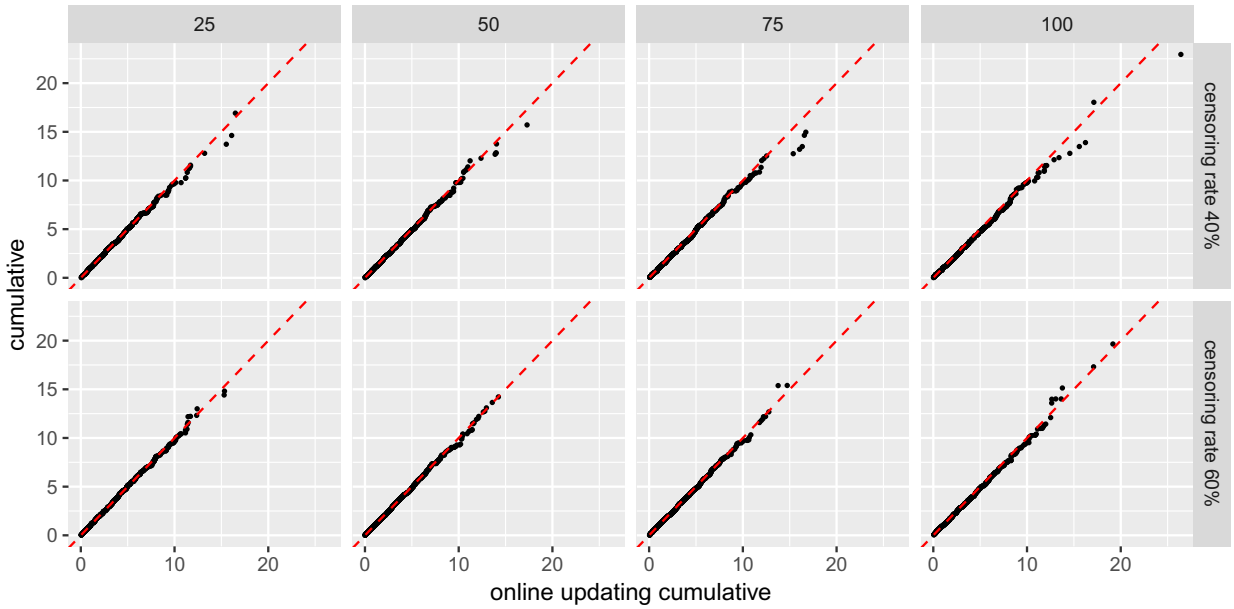


Figure 2. Empirical quantile-quantile plots of the online updating cumulative statistics $T_k(G)$ (x-axis) and $T_{1:k}(G)$ obtained using cumulative data (y-axis) with censoring rate 40% and 60%, taken at block $k \in \{25, 50, 75, 100\}$, both calculated using the KM transformation on event times. This figure appears in color in the electronic version of this article.

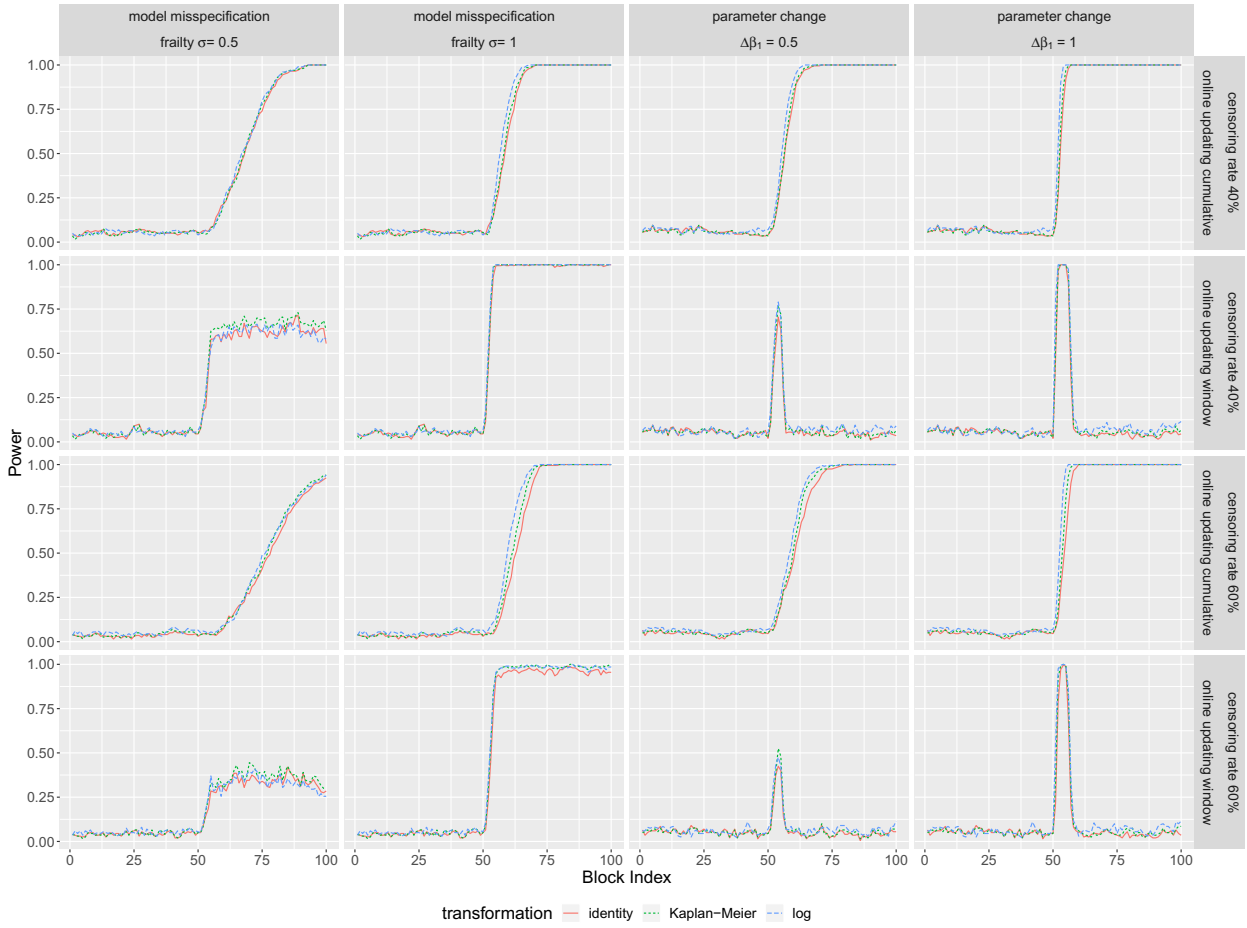


Figure 3. Empirical power (proportion of statistic values greater than $\chi^2_{3,0.95}$) for the online updating cumulative and window tests, calculated at each update using the identity, KM, and log transformations under the alternative hypotheses of model misspecification (left) and parameter change (right) under censoring rate 40% (top) and 60% (bottom). This figure appears in color in the electronic version of this article.

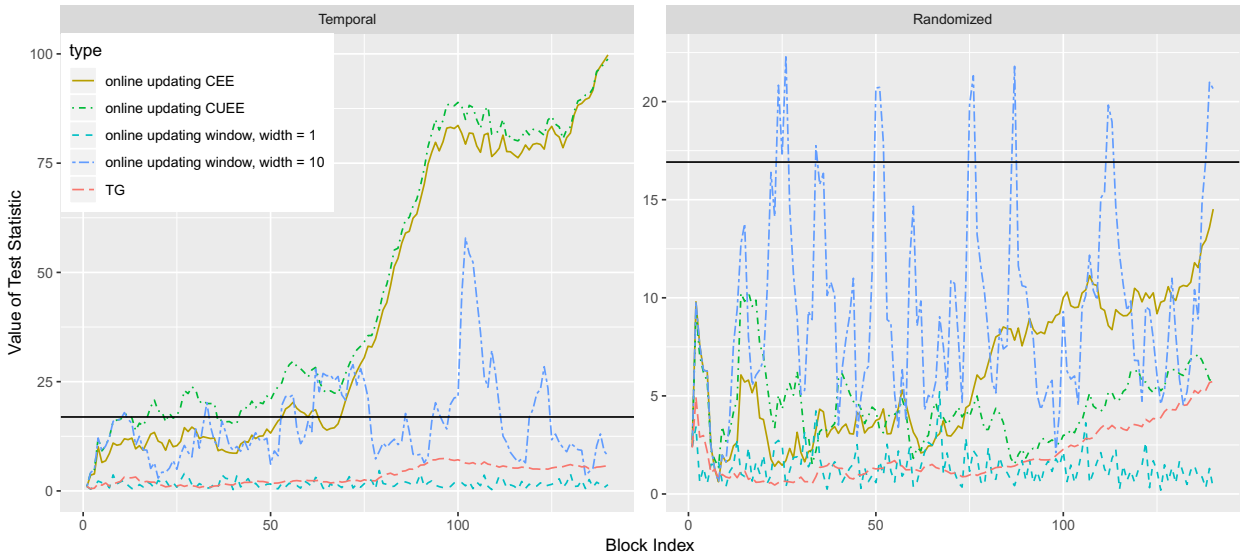


Figure 4. Test statistics for the PH assumption for lymphoma data, using temporally ordered (left) and randomly ordered (right) datasets. This figure appears in color in the electronic version of this article.

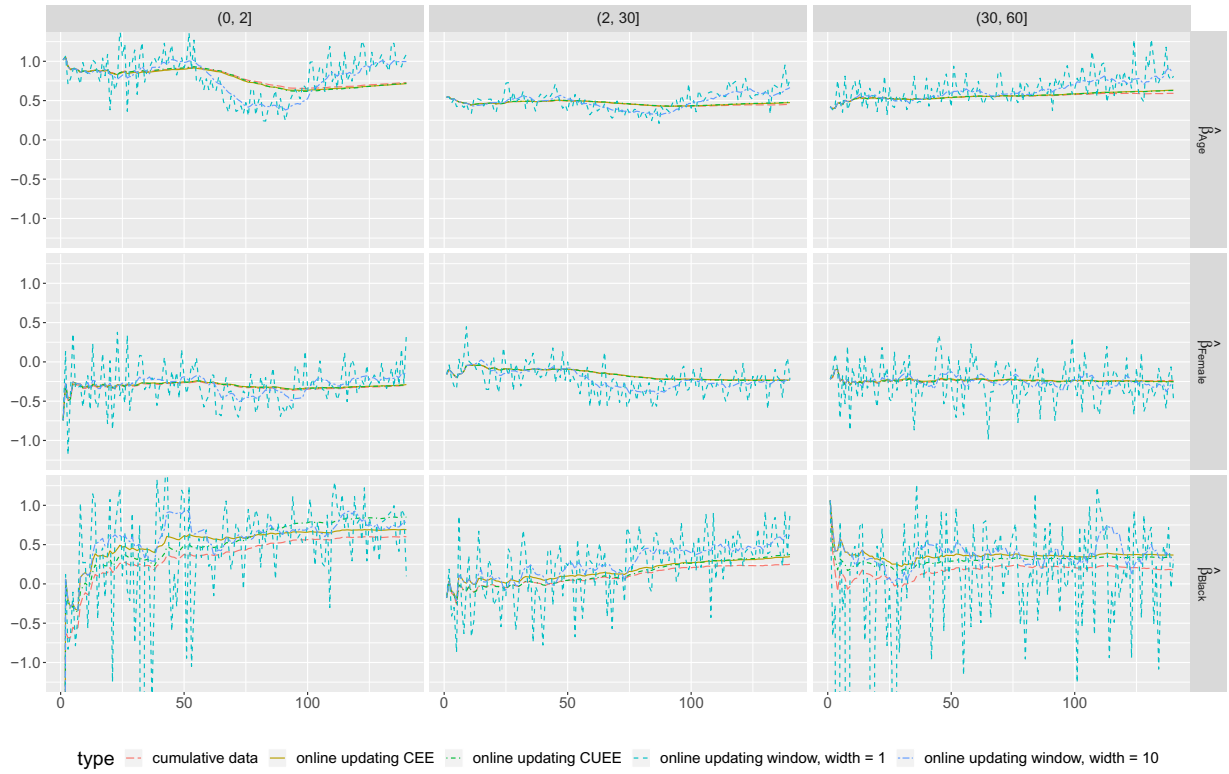


Figure 5. Parameter estimates for Age, Female and Black on intervals $(0, 2]$, $(2, 30]$ and $(30, 60]$, given by different estimating schemes on the temporally ordered lymphoma dataset, plotted against block indices. The decreasing and then increasing trend in the first piece of $\widehat{\beta}_{Age}$ is clear. This figure appears in color in the electronic version of this article.

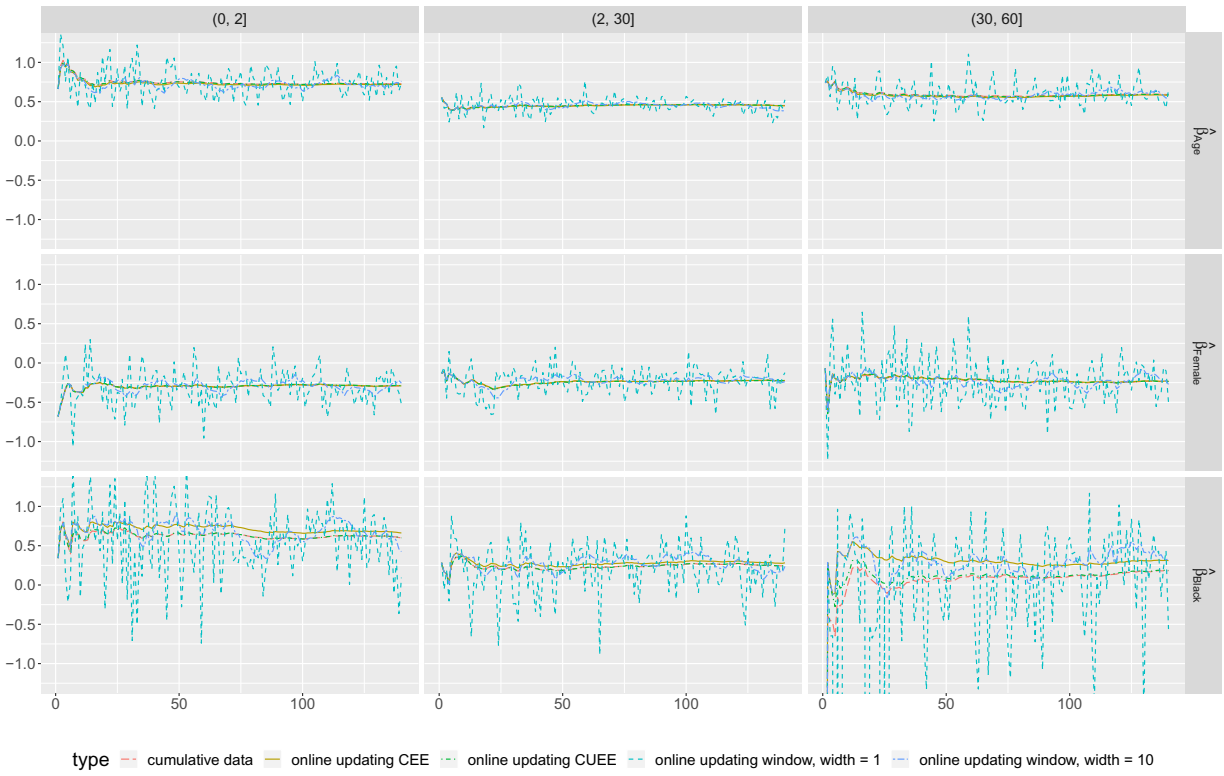


Figure 6. Parameter estimates for Age, Female and Black on intervals $(0, 2]$, $(2, 30]$ and $(30, 60]$, given by different estimating schemes on the randomly ordered lymphoma dataset, plotted against block indices. While the blockwise parameter estimates (window version with $w = 1$) are still volatile, there is no significant increasing or decreasing trend from older blocks to newer blocks. This figure appears in color in the electronic version of this article.

Supporting information for An Online Updating Approach for Testing the Proportional Hazards Assumption with Streams of Survival Data

by

Yishu Xue, HaiYing Wang, Jun Yan, and Elizabeth D. Schifano

Web Appendix A. Proof of Theorem 3.1

The following regularity assumptions are required to establish the asymptotic distribution.

C1 We assume the regularity conditions A-D in Section 2.4 of Andersen and Gill (1982).

C2 The function $g(t)$, $t \in [0, \tau]$, is bounded, where τ is the follow-up time.

C3 Assume that $\{\boldsymbol{X}(t), t \in [0, \tau]\}$ is a bounded Donsker class (Kosorok, 2008).

C4 There exists an $\alpha \in (1/4, 1/2)$ such that for any $\eta > 0$, the subdata estimator $\widehat{\boldsymbol{\beta}}_{n,k}$ satisfies $P(n^{\alpha}\|\widehat{\boldsymbol{\beta}}_{n,k} - \boldsymbol{\beta}_0\| > \eta) \leqslant C_{\eta} n^{2\alpha-1}$, where $C_{\eta} > 0$ is a constant only depending on η.

C5 For each subdata, $\| \sum_{\ell=1}^{d_k} \boldsymbol{G}_{\ell k} \widehat{\boldsymbol{V}}_{\ell k}\| < C_{gv} n \|\widehat{\boldsymbol{\beta}}_{n,k} - \boldsymbol{\beta}_0\|$, or $\| \sum_{\ell=1}^{d_k} \boldsymbol{G}_{\ell k} \check{\boldsymbol{V}}_{\ell k}\| < C_{gv} n \|\check{\boldsymbol{\beta}}_{n,k} - \boldsymbol{\beta}_0\|$, where C_{gv} is a constant that does not depend on k.

The conditions assumed in Section 2.4 of Andersen and Gill (1982) are commonly used in the literature of survival analysis. Since $g(t)$ is user-specified, it is reasonable to assume that it is bounded. Most widely used $g(t)$ functions are bounded if the follow-up time is finite. Condition C3 imposes a constraint on the time dependent covariate. If it is time independent, the condition can be replaced by bounded covariate. Condition C4 is a typical assumption required for online updating method such as in Lin and Xi (2011); Schifano et al. (2016). Condition C5 indicates that $\| \sum_{\ell=1}^{d_k} \boldsymbol{G}_{\ell k} \widehat{\boldsymbol{V}}_{\ell k}\| = O_P(\sqrt{n})$. This condition is typically satisfied in practice. As mentioned in Therneau and Grambsch (2000), $\widehat{\boldsymbol{V}}_{\ell k}$ are often replaced by $\boldsymbol{\mathcal{I}}_{n_k,k}/d_k$ in practice and $\boldsymbol{G}_{\ell k}$ are always centered. Thus, $\sum_{\ell=1}^{d_k} \boldsymbol{G}_{\ell k} \widehat{\boldsymbol{V}}_{\ell k} = 0$ for this scenario.

Proof. If $K = O(n^{\gamma})$, then any $k \leqslant K$ satisfies this condition. Thus, we only need to prove the result for K.

We first consider the case that $\boldsymbol{\mathcal{I}}_{n,k}$ and $\widehat{\boldsymbol{r}}_{\ell k}$ are evaluated at $\widehat{\boldsymbol{\beta}}_{n,k}$. Denote

$$\boldsymbol{\Gamma}_K = \boldsymbol{H}_K^{-1/2} \boldsymbol{Q}_K, \quad \text{where} \quad \boldsymbol{H}_K = \sum_{k=1}^{K} \sum_{\ell=1}^{d_k} \boldsymbol{G}_{\ell k} \widehat{\boldsymbol{V}}_{\ell k} \boldsymbol{G}_{\ell k}. \tag{A.1}$$

To prove the asymptotic chi-square distribution, we only need to show that $\boldsymbol{\Gamma}_K$ converges in distribution to a p-dimensional multivariate standard normal distribution.

We first show that $(nK)^{-1}\boldsymbol{H}_K$ converges in probability to some positive definite matrix. Note that the function $g(t)$ is bounded. Thus, under the conditions A-D in Andersen and Gill (1982), using arguments similar to those used in the proof of Theorem 3.2 (page 1107-1108) of Andersen and Gill (1982), we have that

$$\frac{1}{n}\sum_{\ell=1}^{d_k}\boldsymbol{G}_{\ell k}\widehat{\boldsymbol{V}}_{\ell k}\boldsymbol{G}_{\ell k} \to \int_0^\tau \boldsymbol{G}(t)v(\boldsymbol{\beta}_0,t)\boldsymbol{G}(t)s^{(0)}(\boldsymbol{\beta}_0,t)\lambda_0(t)\mathrm{d}t \equiv \boldsymbol{\Sigma}, \tag{A.2}$$

in probability, where $\boldsymbol{v}(\boldsymbol{\beta},t)$ and $s^{(0)}(\boldsymbol{\beta},t)$ are limits (uniformly in probability) of $\boldsymbol{V}(\boldsymbol{\beta},t)$ and $S^{(0)}(\boldsymbol{\beta},t) = n^{-1}\sum_{i=1}^n Y_i(t)\exp\{\boldsymbol{X}_i(t)^\top\boldsymbol{\beta}\}$, respectively as defined in Conditions A and D in Andersen and Gill (1982).

Since $\{\boldsymbol{X}(t), t\in[0,\tau]\}$ is a bounded Donsker class, $\{Y(t)\exp\{\boldsymbol{X}(t)^\top\boldsymbol{\beta}\}, t\in[0,\tau], \boldsymbol{\beta}\in\mathbb{B}\}$ is also Donsker. A Donsker class is also a Glivenko-Cantelli class, so we have

$$\sup_{t\in[0,\tau],\boldsymbol{\beta}\in\mathbb{B}}\left|\frac{1}{n}\sum_0^n Y_{\ell k}(t)\exp\{\boldsymbol{X}_{\ell k}(t)^\top\boldsymbol{\beta}'\} - s^{(0)}(\boldsymbol{\beta},t)\right| \to 0, \tag{A.3}$$

almost surely, where $\mathbb{B}$ is the compact parameter space. This means that $S^{(0)}(\boldsymbol{\beta},t)$ is uniformly bounded away from 0. As a result, $\frac{1}{n}\sum_{\ell=1}^{d_k}\boldsymbol{G}_{\ell k}\widehat{\boldsymbol{V}}_{\ell k}\boldsymbol{G}_{\ell k}$ is bounded since the covariate $\boldsymbol{X}(t)$ is bounded. Thus, from Theorem 1.3.6 of Serfling (1980), Equation (A.2) implies that

$$E\left(\frac{1}{n}\sum_{\ell=1}^{d_k}\boldsymbol{G}_{\ell k}\widehat{\boldsymbol{V}}_{\ell k}\boldsymbol{G}_{\ell k}\right) \to \boldsymbol{\Sigma}.$$

With this, from Fubini's theorem, we have

$$E\left(\frac{\boldsymbol{H}_K}{nK}\right) = \frac{1}{K}\sum_{k=1}^K E\left(\frac{1}{n}\sum_{\ell=1}^{d_k}\boldsymbol{G}_{\ell k}\widehat{\boldsymbol{V}}_{\ell k}\boldsymbol{G}_{\ell k}\right) \to \boldsymbol{\Sigma}.$$

Thus,

$$\frac{\boldsymbol{H}_K}{nK} \to \boldsymbol{\Sigma}, \tag{A.4}$$

in probability.

Now we examine $\boldsymbol{Q}_K = \sum_{k=1}^K\sum_{\ell=1}^{d_k}\boldsymbol{G}_{\ell k}\widehat{\boldsymbol{r}}_{\ell k}$. For each component of $\widehat{\boldsymbol{r}}_{\ell k}$, $\widehat{r}_{\ell k}^{(i)}$ $(i=1,...,p)$,

the Taylor series expansion yields

$$\widehat{r}_{\ell k}^{(i)} = r_{\ell k}^{(i)} - \boldsymbol{V}_{(i)}(\widehat{\boldsymbol{\beta}}_{n,k}^{(i*)}, t_\ell)(\widehat{\boldsymbol{\beta}}_{n,k} - \boldsymbol{\beta}_0),$$

where $\boldsymbol{V}_{(i)}(\widehat{\boldsymbol{\beta}}_{n,k}^{(i*)}, t_\ell)$ is the ith row of $\boldsymbol{V}(\widehat{\boldsymbol{\beta}}_{n,k}^{(i*)}, t_\ell)$, and $\widehat{\boldsymbol{\beta}}_{n,k}^{(i*)}$ is on the line segment between $\widehat{\boldsymbol{\beta}}_{n,k}$ and $\boldsymbol{\beta}_0$. If $\boldsymbol{V}(\widehat{\boldsymbol{\beta}}_{n,k}^{*}, t_\ell)$ is the matrix whose rows are $\boldsymbol{V}_{(i)}(\widehat{\boldsymbol{\beta}}_{n,k}^{(i*)}, t_\ell)$, $i = 1, ..., p$, then we have

$$\widehat{\boldsymbol{r}}_{\ell k} = \boldsymbol{r}_{\ell k} - \boldsymbol{V}(\widehat{\boldsymbol{\beta}}_{n,k}^{*}, t_\ell)(\widehat{\boldsymbol{\beta}}_{n,k} - \boldsymbol{\beta}_0).$$

Thus

$$\begin{aligned} \boldsymbol{Q}_K &= \sum_{k=1}^{K}\sum_{\ell=1}^{d_k} \boldsymbol{G}_{\ell k}\widehat{\boldsymbol{r}}_{\ell k} \\ &= \sum_{k=1}^{K}\sum_{\ell=1}^{d_k} \boldsymbol{G}_{\ell k} r_{\ell k} - \sum_{k=1}^{K}\sum_{\ell=1}^{d_k} \boldsymbol{G}_{\ell k}\boldsymbol{V}(\widehat{\boldsymbol{\beta}}_{n,k}^{*}, t_\ell)(\widehat{\boldsymbol{\beta}}_{n,k} - \boldsymbol{\beta}_0) \equiv \boldsymbol{\Delta}_1 - \boldsymbol{\Delta}_2. \end{aligned} \tag{A.5}$$

Note that $\boldsymbol{\Delta}_1$ is a weighted score function for the full data log partial likelihood, and the weights are bounded. Thus, using arguments similar to the those used in the proof of Theorem 3.2 (pages 1106–1107) of Andersen and Gill (1982), we know that

$$\frac{\boldsymbol{\Delta}_1}{\sqrt{nK}} \to N(0, \boldsymbol{\Sigma}), \tag{A.6}$$

in distribution. Now we show that

$$\frac{\boldsymbol{\Delta}_2}{\sqrt{nK}} = o_P(1). \tag{A.7}$$

Note that for each k, $\| \sum_{\ell=1}^{d_k} \boldsymbol{G}_{\ell k}\widehat{\boldsymbol{V}}_{\ell k}\| < C_{gv} n \|\widehat{\boldsymbol{\beta}}_{n,k} - \boldsymbol{\beta}_0\|$. Thus,

$$\begin{aligned} \|\boldsymbol{\Delta}_2\| \leqslant & \sum_{k=1}^{K}\left\|\sum_{\ell=1}^{d_k} \boldsymbol{G}_{\ell k}\{\boldsymbol{V}(\widehat{\boldsymbol{\beta}}_{n,k}^{*}, t_\ell) - \boldsymbol{V}(\widehat{\boldsymbol{\beta}}_{n,k}, t_\ell)\}(\widehat{\boldsymbol{\beta}}_{n,k} - \boldsymbol{\beta}_0)\right\| \\ & + \sum_{k=1}^{K}\left\|\sum_{\ell=1}^{d_k} \boldsymbol{G}_{\ell k}\boldsymbol{V}(\widehat{\boldsymbol{\beta}}_{n,k}, t_\ell)(\widehat{\boldsymbol{\beta}}_{n,k} - \boldsymbol{\beta}_0)\right\| \\ \leqslant & C_g \sum_{k=1}^{K}\sum_{\ell=1}^{d_k} \|\boldsymbol{V}(\widehat{\boldsymbol{\beta}}_{n,k}^{*}, t_\ell) - \boldsymbol{V}(\widehat{\boldsymbol{\beta}}_{n,k}, t_\ell)\| \|\widehat{\boldsymbol{\beta}}_{n,k} - \boldsymbol{\beta}_0\| + C_{gv} n \sum_{k=1}^{K} \|\widehat{\boldsymbol{\beta}}_{n,k} - \boldsymbol{\beta}_0\|^2, \end{aligned} \tag{A.8}$$

where C_g is a constant that bounds $\boldsymbol{G}(t)$ from above.

For the i_1i_2th element of $V(\widehat{\boldsymbol{\beta}}^*_{n,k}, t_\ell) - V(\widehat{\boldsymbol{\beta}}_{n,k}, t_\ell)$,

$$\boldsymbol{V}_{(i_1i_2)}(\widehat{\boldsymbol{\beta}}^*_{n,k}, t_\ell) - \boldsymbol{V}_{(i_1i_2)}(\widehat{\boldsymbol{\beta}}_{n,k}, t_\ell) = \frac{\partial \boldsymbol{V}_{(i_1i_2)}(\widehat{\boldsymbol{\beta}}^{**}_{n,k}, t_\ell)}{\partial \boldsymbol{\beta}}(\widehat{\boldsymbol{\beta}}^*_{n,k} - \widehat{\boldsymbol{\beta}}_{n,k}),$$

where $\widehat{\boldsymbol{\beta}}^{**}_{n,k}$ is on the line segment between $\widehat{\boldsymbol{\beta}}^*_{n,k}$ and $\widehat{\boldsymbol{\beta}}_{n,k}$. From (A.3) and the fact that $\boldsymbol{X}(t)$ is bounded, we know that $\partial \boldsymbol{V}_{(i_1i_2)}(\widehat{\boldsymbol{\beta}}^{**}_{n,k}, t_\ell)/\partial \boldsymbol{\beta}$ is uniformly bounded. Let M be a constant that bounds its elements. Since $\widehat{\boldsymbol{\beta}}^{**}_{n,k}$ and $\widehat{\boldsymbol{\beta}}^*_{n,k}$ are between $\widehat{\boldsymbol{\beta}}_{n,k}$ and $\boldsymbol{\beta}_0$, we have

$$|\boldsymbol{V}_{(i_1i_2)}(\widehat{\boldsymbol{\beta}}^*_{n,k}, t_\ell) - \boldsymbol{V}_{(i_1i_2)}(\widehat{\boldsymbol{\beta}}_{n,k}, t_\ell)| \leqslant M\|\widehat{\boldsymbol{\beta}}_{n,k} - \boldsymbol{\beta}_0\|. \tag{A.9}$$

Combining (A.8) and (A.9), we have

$$\|\boldsymbol{\Delta}_2\| \leqslant Cn\sum_{k=1}^{K}\|\widehat{\boldsymbol{\beta}}_{n,k} - \boldsymbol{\beta}_0\|^2, \tag{A.10}$$

where $C = C_gM + C_{gv}$. Since $K = O(n^\gamma)$, there exist a constant, say C_1^2, such that $K < C_1^2 n^\gamma$. From (A.10), for any $\epsilon > 0$,

$$\begin{aligned}
P\left(\|\boldsymbol{\Delta}_2\| > \sqrt{nK}\epsilon\right) &\leqslant P\left(\frac{1}{K}\sum_{k=1}^{K}\left\|\widehat{\boldsymbol{\beta}}_{n,k} - \boldsymbol{\beta}_0\right\|^2 > \frac{\epsilon}{C\sqrt{nK}}\right) \\
&\leqslant \sum_{k=1}^{K} P\left(\left\|\widehat{\boldsymbol{\beta}}_{n,k} - \boldsymbol{\beta}_0\right\|^2 > \frac{\epsilon}{C\sqrt{nK}}\right) \\
&\leqslant \sum_{k=1}^{K} P\left(\sqrt{nn^\gamma}\left\|\widehat{\boldsymbol{\beta}}_{n,k} - \boldsymbol{\beta}_0\right\|^2 > \frac{\epsilon}{CC_1}\right) \\
&= \sum_{k=1}^{K} P\left(n^{(1+\gamma)/4}\left\|\widehat{\boldsymbol{\beta}}_{n,k} - \boldsymbol{\beta}_0\right\| > \sqrt{\frac{\epsilon}{CC_1}}\right) \\
&\leqslant \sum_{k=1}^{K} P\left(n^{\alpha}\left\|\widehat{\boldsymbol{\beta}}_{n,k} - \boldsymbol{\beta}_0\right\| > \sqrt{\frac{\epsilon}{CC_1}}\right) \\
&\leqslant \sum_{k=1}^{K} C_\eta n^{2\alpha-1} = C_\eta K n^{2\alpha-1} = O(n^{\gamma+2\alpha-1}) = o(1).
\end{aligned}$$

Here, the last inequality is from condition C4; the second last inequality is because $\gamma < 4\alpha - 1$; and the last step is because $\gamma < 1 - 2\alpha$. This proves (A.7). The proof finishes by combining (A.1), (A.4), (A.5), (A.6), (A.7), and Slutsky's theorem.

Now we consider the case when $\boldsymbol{\mathcal{I}}_{n,k}$ and $\widehat{\boldsymbol{r}}_{\ell k}$ are evaluated at $\breve{\boldsymbol{\beta}}_{n,k}$. Under Condition C1

and C4, the requirements of (C4') and (C6) in Lemma E.2 of Schifano et al. (2016) are satisfied. Thus, the condition described in C4 for $\widehat{\boldsymbol{\beta}}_{n,k}$ is also valid for $\breve{\boldsymbol{\beta}}_{n,k}$. With this result, the proof is similar to the case when $\boldsymbol{\mathcal{I}}_{n,k}$ and $\widehat{\boldsymbol{r}}_{\ell k}$ are evaluated at $\widehat{\boldsymbol{\beta}}_{n,k}$.

Web Appendix B. Additional Simulation Results

B.1 Computing Time Comparison

In this section, we present the computation time for the standard test $T(\boldsymbol{G})$ and the online updating cumulative statistic $T_k(\boldsymbol{G})$, for both the CEE- and CUEE-based versions. In this comparison, data generation and data loading time is not recorded, but only the computation time. Survival data streams using the setting of Section 4.1 with $\varepsilon = 0.1$ are generated. The size of the stream, N, is such that $N \in \{100000, 200000, 300000, 400000, 500000\}$, and each stream is partitioned into 100 equally sized blocks, such that $n_k \in \{1000, 2000, 3000, 4000, 5000\}$ for $k = 1, \ldots, 100$. For each stream, the time it takes to calculate the maximum partial likelihood estimate of $\boldsymbol{\beta}$ and the diagnostic statistic $T(\boldsymbol{G})$ are recorded, as well as the time it takes to obtain $T_k(\boldsymbol{G})$, $\widehat{\boldsymbol{\beta}}_k$ and $\widetilde{\boldsymbol{\beta}}_k$ for $k = 1, \ldots, 100$. The results are obtained for 100 replicates of simulation performed with Intel® Core(TM) i7-8850H CPU @2.60GHz, and we illustrate the computing time in Web Figure 1. It is rather apparent that the standard test is far more time-consuming than both versions of the proposed online updating cumulative test, and the disparity increases with the size of the data stream. The CUEE-based $T_k(\boldsymbol{G})$ is slightly slower than the CEE-based $T_k(\boldsymbol{G})$, but the difference is minor. Note that $T(\boldsymbol{G})$ is only computed once, at the end of each stream. If we want to obtain a new $T(\boldsymbol{G})$ on cumulative data upon the arrival of each new block, like we can do with $T_k(\boldsymbol{G})$, the contrast of computing time would be even more significant.

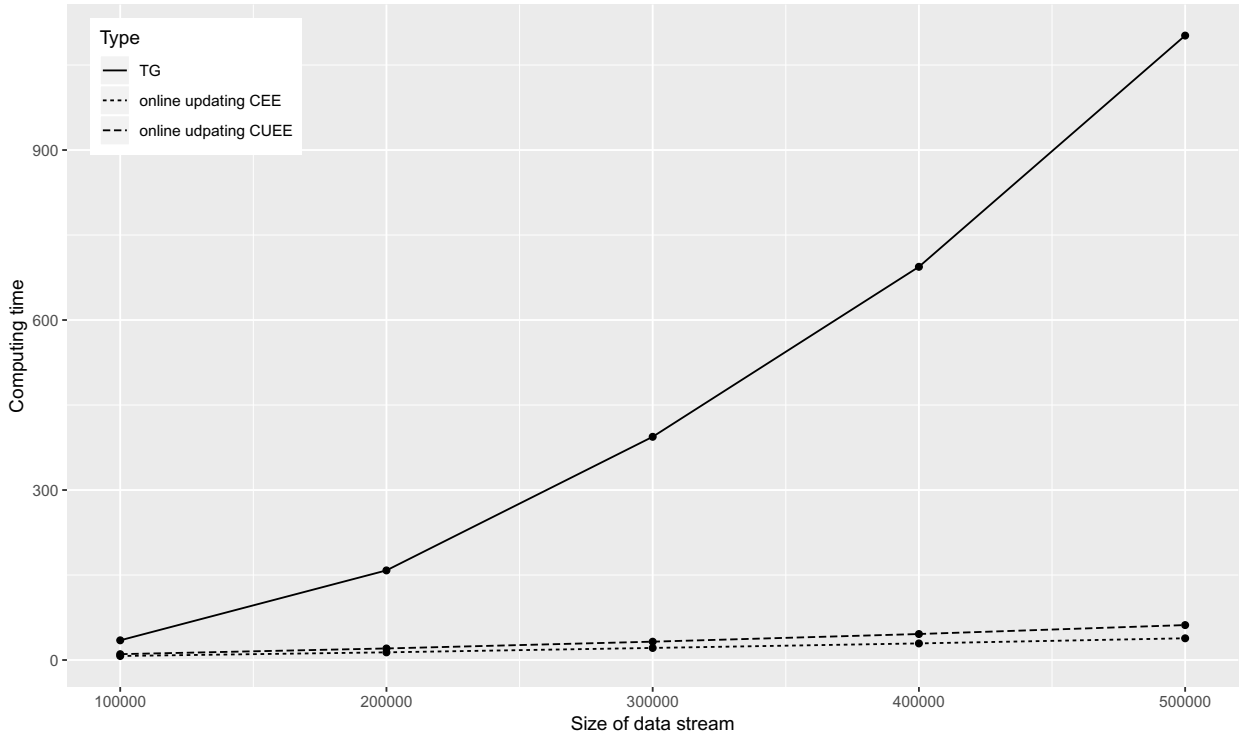
Next we present a brief time complexity study. To compute $T(\boldsymbol{G})$ on a dataset with N observations and d events, we first need to evaluate the log partial likelihood (3). The summation inside the logarithm has $O(N)$ complexity, while the outer integral is indeed a summation

over d individual event times, which requires computing the component inside the square brackets for d times. Therefore evaluation of the partial likelihood has $O(Nd)$ complexity. Assuming that d is roughly of the same order as N, this is equivalent to $O(N^2)$ complexity. Calculation of the Schoenfeld residuals, similarly, is roughly $O(N^2)$. Other procedures in Equation (6) include multiplication of $1 \times d$, $d \times p$, and $p \times p$ matrices, and the inversion of $p \times p$ matrices, and the time complexity is capped at $O(dp + p^3 + p^2)$, which is dominated by $O(N^2)$ when the number of events is much larger than the dimension of covariate space and therefore ignored.

The online updating approach breaks the dataset into K blocks. For simplicity let us assume the block sizes are all equal to N/K, then evaluating the partial likelihood, together with calculation of the Schoenfeld residuals, has $O(N^2/K^2)$ complexity, therefore doing so for all K blocks will require $O(N^2/K)$ time. This indicates that the speed of online updating is inversely proportional to the number of blocks that a dataset is partitioned into. Note, however, that K needs to satisfy the regularity condition in Theorem 3.1.

B.2 Memory Usage Assessment

We present a study on memory usage of our proposed online updating statistics. A big dataset was simulated using the parameter setting in Section 4.1 with $\boldsymbol{\beta}_0 = (0.67, -0.26, 0.36)$ and $\lambda_0(t) = 0.018$, which contains $N = 200$ million observations. The size of the simulated dataset, when written into a csv file is 7.65 GB. Using the **bigmemory** package (Kane et al., 2013), a description file is created, which contains references to the same dataset but converted to a C++ object, stored on the hard drive. The description file can be loaded after it is created to allow access of the corresponding data from within R, without having to load the entire dataset into the memory. All studies were performed under single-core mode on the same laptop as in Web Appendix B.1. The total memory available on this laptop is 32 GB. The **profvis** package (Chang and Luraschi, 2018) was used to track the



Web Figure 1: Plot of average computing time versus size of data stream for 100 replicates of simulation for $T(\boldsymbol{G})$ and two versions of $T_k(\boldsymbol{G})$.

memory usage and running time. The block size is chosen to be $n_k = 2000$, resulting in 10,000 blocks in total. Creation of the description profile takes 407.5 seconds, and the cumulative memory usage was 16,785.2 MB. Next, the online updating CUEE-based $T_k(\boldsymbol{G})$ was calculated for the 10,000 blocks. At each update, memory was first allocated and then de-allocated after the blockwise summary statistics were obtained and the data block was removed. The cumulative memory allocation for loading the description file and performing online updating diagnostics was 43,318.2 MB, and the cumulative memory de-allocation was 43,297.4 MB, which indicates that on average, each update requires slightly more than 4 MB memory. The entire data loading, model estimation and diagnostic process took 1,048 seconds.

As a comparison, we also tried to read the entire dataset into R's workspace and perform the standard analysis on the whole data. The read.csv() procedure was attempted first, but it did not finish after running for more than an hour, and was finally aborted due to memory

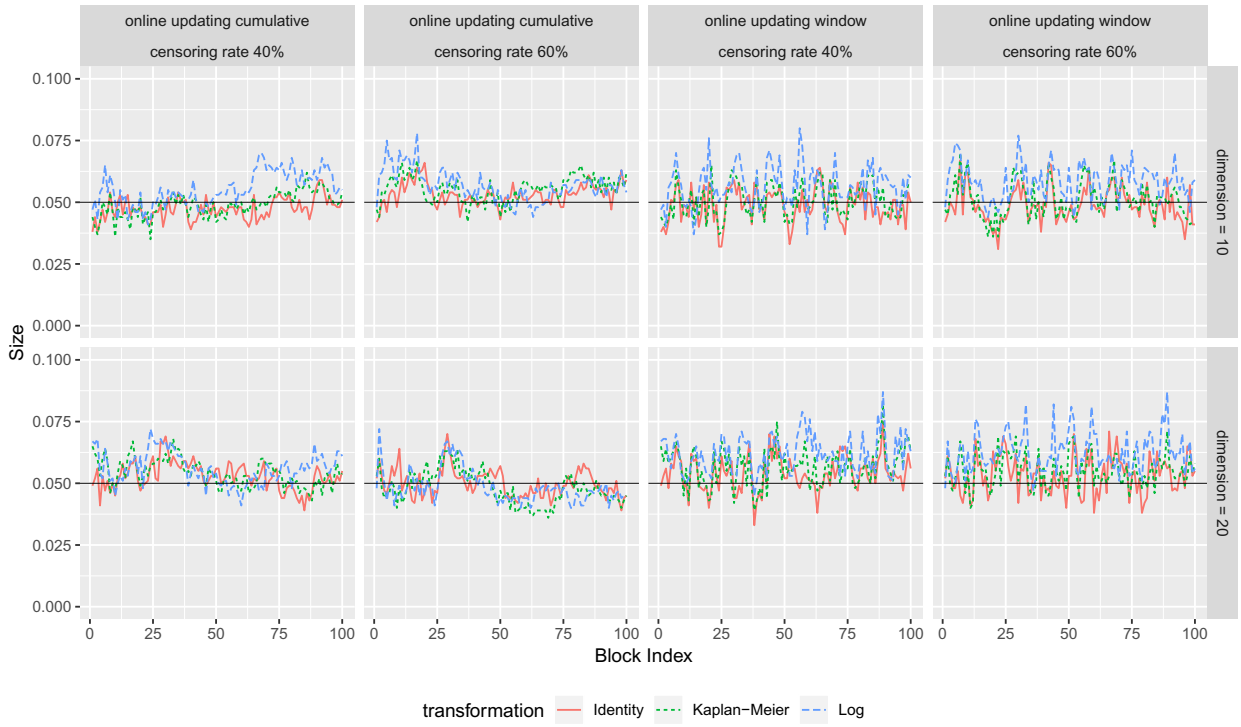
insufficiency. The fread() function in the **data.table** package (Dowle and Srinivasan, 2019), which has been known for fast reading of big datasets, was also attempted. The data reading process itself took 4,325 seconds. After the data was loaded, however, even simple operations (e.g., obtaining summaries of covariate distributions) could not be completed, and the fitting of the Cox model was not attempted.

B.3 Sizes under Moderate Dimensions

We present additional simulation results for the size of the proposed test statistics for $p \in \{10, 20\}$. For each setting, there are $p/2$ continuous covariates, generated i.i.d. from $N(0, 1)$, and the remaining $p/2$ covariates are binary, generated i.i.d. from Bernoulli(0.5). The vectors of coefficients are chosen as $\boldsymbol{\beta}_{10} = (0.7, -0.5, 0.8, 0.3, 0.1, -0.4, -0.9, -0.2, -0.3, 0.4)^\top$, $\boldsymbol{\beta}_{20} = (\boldsymbol{\beta}_{10}^\top, \boldsymbol{\beta}_{10}^\top)^\top$. The baseline hazards are set to, respectively, 0.032 and 0.015, with the weights at $\langle 60 \rangle$ being $(0.9, 0.1)$ to produce the desired censoring rates of approximately 40% and 60%. For each scenario considered, 1,000 replicates of simulation are performed. It can be seen from Web Figure 2 that both versions of statistic hold their sizes under the null hypothesis, under both dimensions, although the log transformation is not recommended.

B.4 Sizes with Piecewise Constant Coefficients

Because our initial analysis of the SEER lymphoma data suggested a Cox model with time-varying coefficients that could be approximated by a piecewise constant function of time, we checked the size of the proposed test in a simulation study with a Cox model having a similar structure. The function survSplit() from R package **survival** (Therneau, 2015) facilitates the fitting of Cox models for these piecewise-constant time-varying coefficients with the use of tgroup as described in Section 5 and further detailed in Therneau et al. (2018). As an illustration, we used the **reda** package (Wang et al., 2017) to simulate survival data with the three covariates as in Section 4, but the coefficients are now piecewise constant. On the interval $(0, 12]$, $\boldsymbol{\beta}_0 = (0.7, -0.26, 0.36)$, and on the interval $(12, 60]$, $\boldsymbol{\beta}_0 = (0.6, -0.4, 0.46)$.



Web Figure 2: Size for the proposed test statistics when $p = 10$ and 20.

Web Table 1: Size of $T_k(\boldsymbol{G})$ for models with piecewise constant coefficients based on 1,000 replicates.

Censoring Rate	Transformation	Size
40%	Kaplan–Meier	0.067
	Identity	0.043
	Logarithm	0.156
60%	Kaplan–Meier	0.039
	Identity	0.033
	Logarithm	0.094

The same censoring schemes as in Section 4 have been used and produced censoring rates of approximately 40% and 60%. Function survSplit() was applied with breaking point 12. The online updating cumulative statistic $T_k(\boldsymbol{G})$ evaluated at the CUEE was compared against critical value $\chi^2_{0.95,6}$ to make the decision. The empirical sizes from the three transformations are summarized in Web Table 1.

For both censoring rates, it can be seen that the empirical type I error rate is appropriately controlled around its nominal level of 0.05 when the Kaplan–Meier or identity transforma-

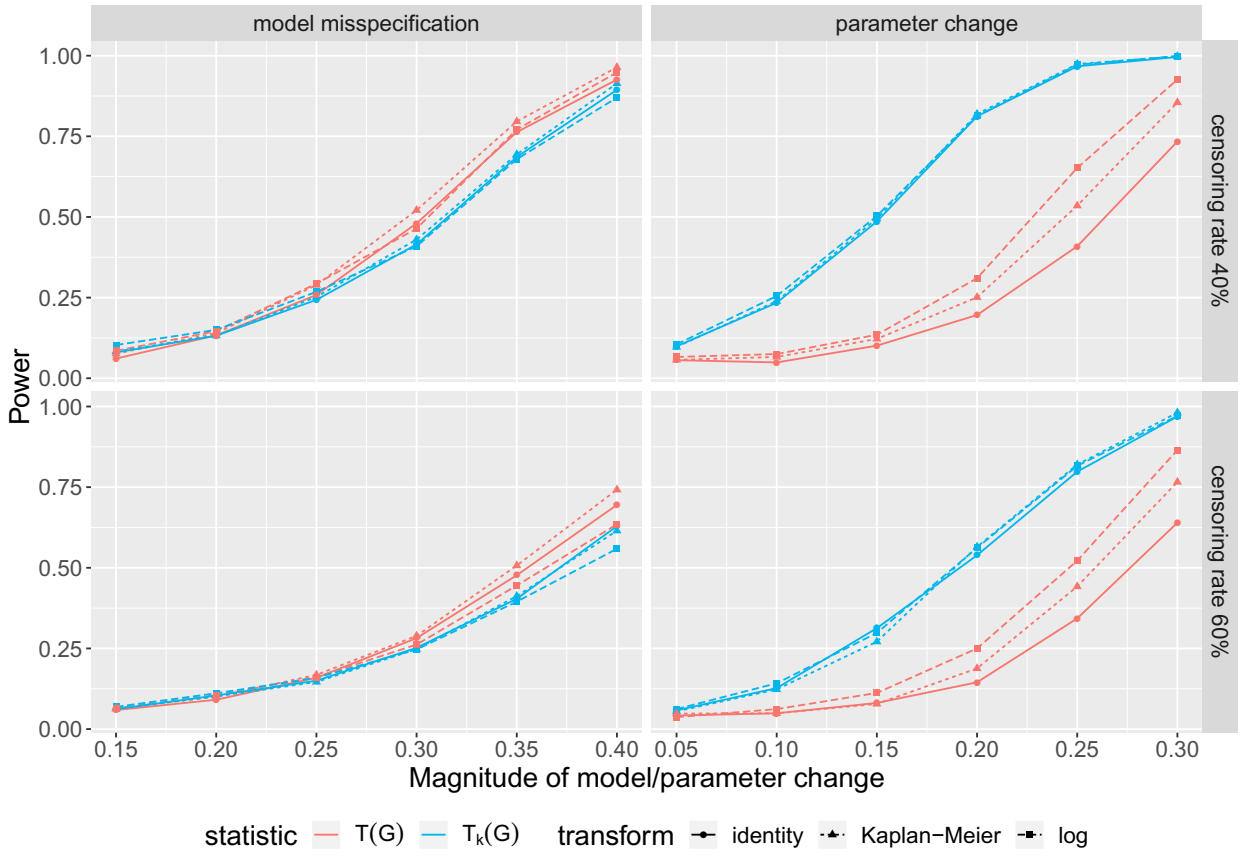
tions are used. The logarithm transformation does not maintain its size well, which is similar to the instability we observed in Figure 1 and Web Figure 2, and is again not recommended.

B.5 Additional Power Comparison

In this section, we present additional simulation results to compare the power of the cumulative version online updating statistic $T_k(\boldsymbol{G})$, and the power of $T(\boldsymbol{G})$ in Grambsch and Therneau (1994), at the end of each data stream, for different types and magnitudes of violations. The simulation setting in Section 4 yielded power of almost 1 for both $T(\boldsymbol{G})$ and $T_k(\boldsymbol{G})$. Therefore we choose to use smaller magnitudes of change in conducting the power comparison. For the model misspecification scenario, we choose $\sigma \in \{0.15, 0.20, 0.25, 0.30, 0.35, 0.40\}$. For each σ, 1000 replicates of simulation are performed, and the power is calculated in the end of the data stream in each replicate for both $T(\boldsymbol{G})$ and $T_k(\boldsymbol{G})$. Similarly for the parameter change scenario, for $\Delta\beta_1 \in \{0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}$, the power for 1000 replicates of simulation is also calculated. All three transformations are assessed under both the low and high censoring rates. We plot the powers against the magnitudes of model/parameter change in Web Figure 3.

It can be seen that, when the violation is due to a model change to frailty, both versions have relatively low power when the frailty standard deviation is small. At $\sigma = 0.40$, however, both $T(\boldsymbol{G})$ and $T_k(\boldsymbol{G})$ identify the violation with quite high power. The performance of $T_k(\boldsymbol{G})$ is not better than, but still comparable to, the performance of $T(\boldsymbol{G})$. Note that both statistics have higher power for the same change at 40% censoring level than at 60% censoring level.

When the violation is due to a change in covariate effects, however, our proposed online updating cumulative statistic $T_k(\boldsymbol{G})$ has significantly higher power than $T(\boldsymbol{G})$. While both statistics have small power at $\Delta\beta_1 = 0.05$, when $\Delta\beta_1$ increases, the power of $T_k(\boldsymbol{G})$ increases faster than the power of $T(\boldsymbol{G})$, and the difference in powers can be as large as nearly 0.5.



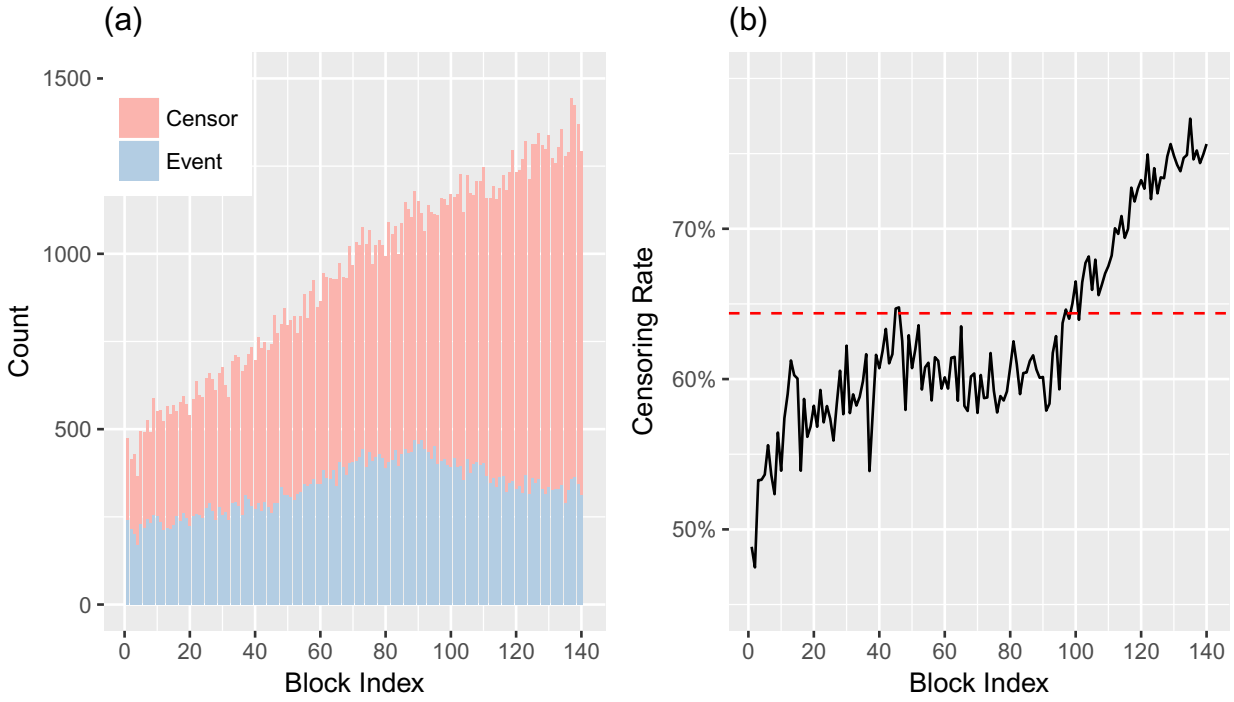
Web Figure 3: Powers of $T(\boldsymbol{G})$ and $T_k(\boldsymbol{G})$ calculated at the end of the data stream when a violation occurs at the 51st block in each stream, plotted against the magnitude of violations. For the model misspecification scenario, the x axis denotes the frailty standard deviation, σ; for the parameter change scenario, the x axis denotes the change in β_1, i.e., $\Delta\beta_1$.

Web Appendix C. Additional Analysis of the SEER Lymphoma Data

C.1 Sample Size and Censoring Rate

Observations in SEER lymphoma data were first ordered by their time of diagnosis. Next, they were grouped by quarter of a year into 140 blocks. The average sample size per block is approximately 943.

Web Figure 4(a) is the stacked bar plot of censors and events in each block. It can be seen that, as a consequence of population increase, the total number of diagnoses per block increases with time. Advancements in medicine, however, helped more recently diagnosed



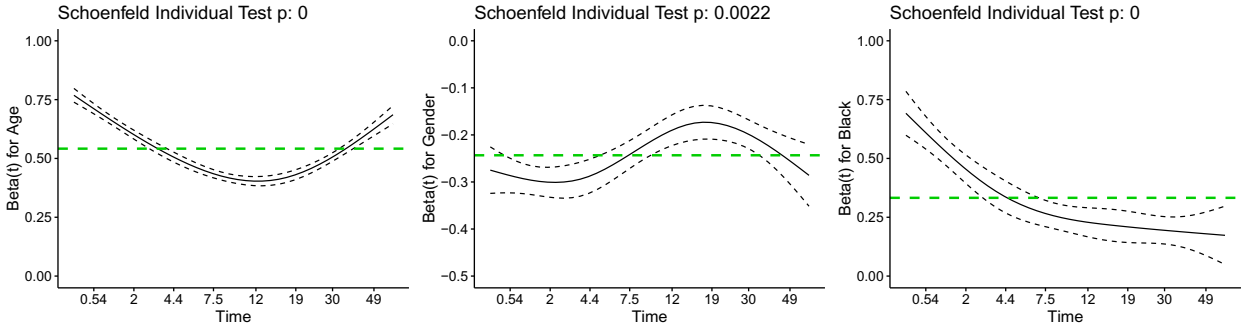
Web Figure 4: Sample size and censoring rate in blocks of SEER lymphoma data.

patients survive more than 5 years, which caused the censoring rate to be fast increasing after year 1995, as shown in Web Figure 4(b).

C.2 Time-Varying Coefficients

Web Figure 5 was obtained by calling `plot.cox.zph()` on the initial model, which enabled us to check if the parameters were time-varying. Notice that the x-axis (event time) will be transformed using the same Kaplan–Meier method as in calculation of the $T(\boldsymbol{G})$ statistic. All three parameters are clearly time-varying. Therefore, it is reasonable to consider using a more flexible model, with the parameters being piecewise constant with different values on different, disjoint intervals of survival time.

As a comparison, we plot in Web Figure 6 the time-varying pattern of parameters in the revised model. In contrast to Web Figure 5, the parameter estimates are much more stable as the confidence band of each parameter estimate at different times contain its entire data estimate for almost the whole time range.



Web Figure 5: Time-varying pattern of the parameters for `Age`, `Gender` and `Black` in the initial model, with parameter estimates from the entire data overlaid in green.

C.3 A Permutation Test

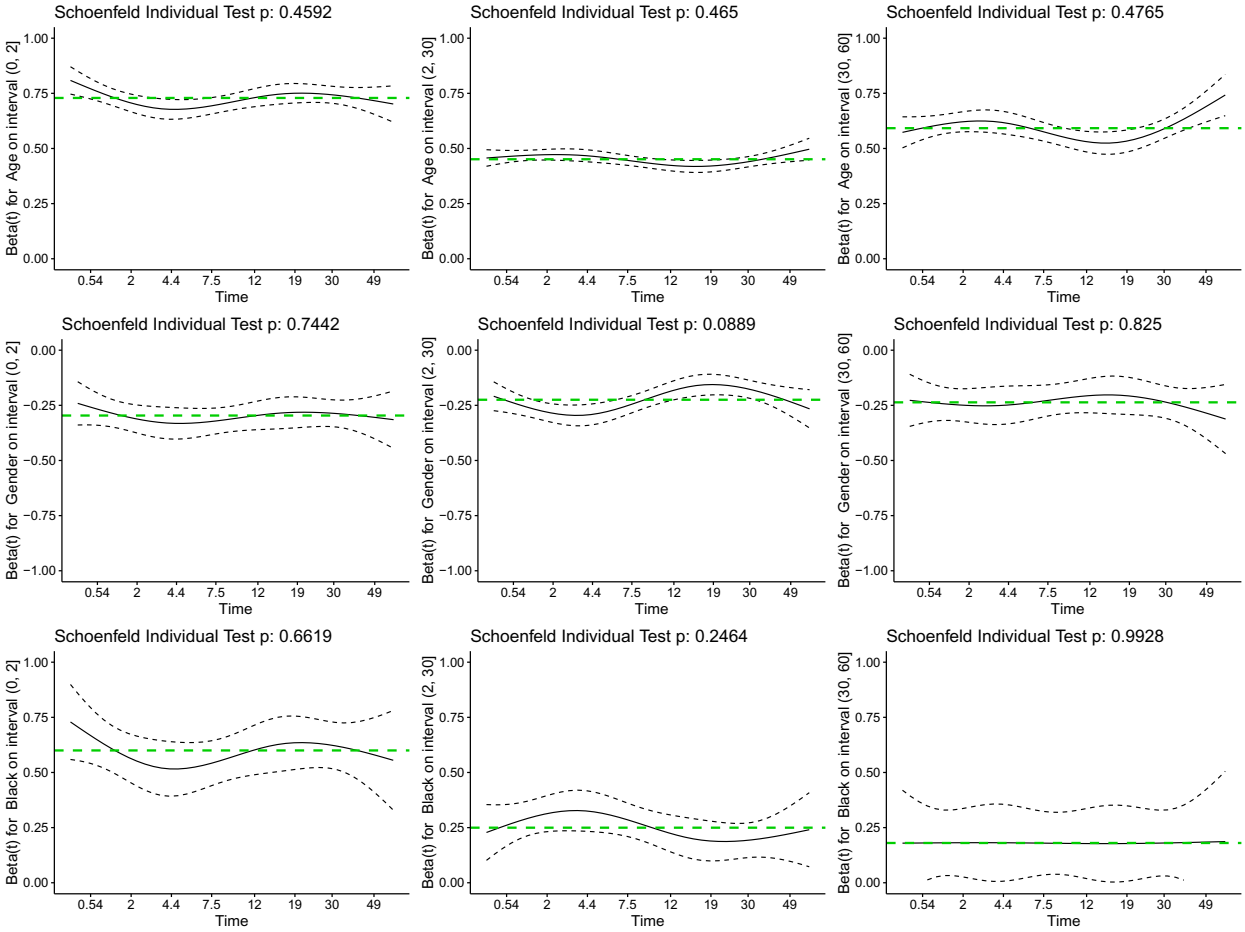
To confirm that the temporal change in parameter contributed to the highly significant online updating cumulative test statistic, we randomly permuted the order of the observations in the original dataset 1,000 times using the same block size as the temporal data. For each permutation, the same techniques and cut-off values were applied to allow the parameters to be piecewise constant over disjoint intervals of survival time.

While there is no guarantee that each permutation of the data produces $\widehat{\boldsymbol{\beta}}$ that is stable between blocks without obvious trends, the online updating cumulative statistics based on permutations of the dataset have a certain distribution. Web Figure 7 is the histogram of online updating cumulative statistics obtained at the final block for 1,000 such permutations. The empirical p-value based on these 1,000 permutations is 0.016, indicating that the particular order of blocks in the original temporally ordered data does contribute to the large values of the online updating test statistics.

References

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *The Annals of Statistics* **10,** 1100–1120.

Chang, W. and Luraschi, J. (2018). *profvis: Interactive Visualizations for Profiling R Code.*



Web Figure 6: Time-varying pattern of the parameters for `Age`, `Gender` and `Black` in the revised model on three disjoint intervals of survival time, with parameter estimates from the entire data overlaid in green.

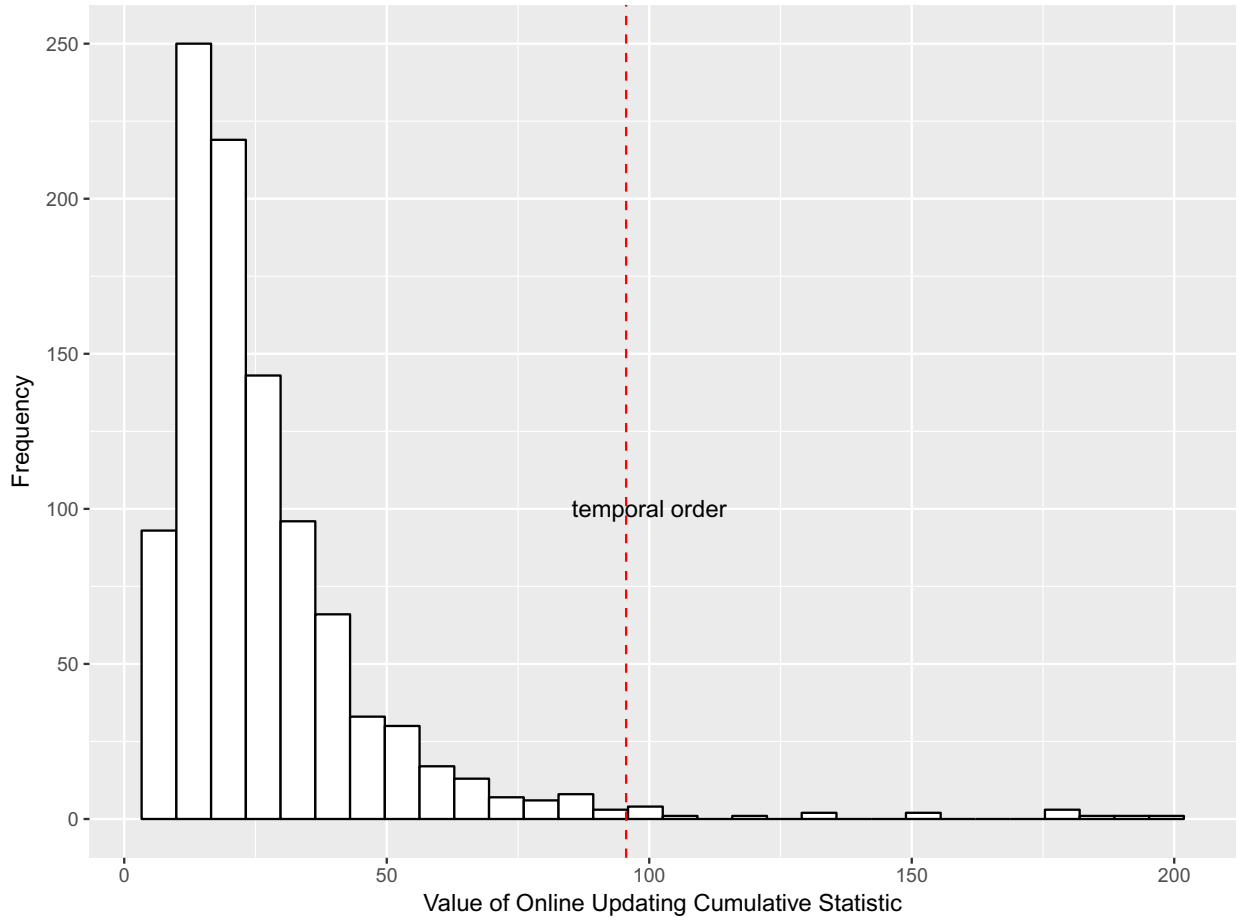
R package version 0.3.5.

Dowle, M. and Srinivasan, A. (2019). *data.table: Extension of 'data.frame'*. R package version 1.12.2.

Grambsch, P. M. and Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* **81,** 515–526.

Kane, M., Emerson, J. W., and Weston, S. (2013). Scalable strategies for computing with massive data. *Journal of Statistical Software* **55,** 1–19.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer.



Web Figure 7: Histogram of online updating cumulative statistic obtained at the final block for 1,000 permutations of the original data, with observed test statistic value for the original data overlaid in red.

Lin, N. and Xi, R. (2011). Aggregated estimating equation estimation. *Statistics and its Interface* **4,** 73–83.

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

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, New York.

Therneau, T., Crowson, C., and Atkinson, E. (2018). Using time dependent covariates and time dependent coefficients in the Cox model.

Therneau, T. M. (2015). *A Package for Survival Analysis in S*. version 2.38.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag Inc, Berlin; New York.

Wang, W., Fu, H., and Yan, J. (2017). *reda: Recurrent Event Data Analysis*. R package version 0.4.1.