4-9-2019

Diagnostic Methods for Big Survival Data

Yishu Xue
University of Connecticut - Storrs, yishu.xue@uconn.edu

Diagnostic Methods for Big Survival Data

Yishu Xue, Ph.D.
University of Connecticut, 2019

ABSTRACT

While studies of the proportional hazards model for big survival data mainly focus on speeding up computation and selecting features from a huge number of covariates, verifying the crucial assumption of proportional hazards (PH) has not been tackled for big data when the data size exceeds a computer's memory. This dissertation summarizes methodological developments in statistics that address the diagnostics of the PH model, including the PH assumption, functional form, and outlying and/or influential observations. Specifically, an online updating approach with minimal storage requirement that updates the standard test statistic for the PH assumption in an online fashion is proposed. The test and its variant based on most recent data blocks maintain their sizes when the PH assumption holds, and have substantial power when it is violated in different ways. Attention has also been paid to the baseline hazard function of the PH model. Nonparametric methods to compare cumulative baseline hazard curves using profile monitoring techniques, and their combination with parametric methods to detect heterogeneity in data blocks, are presented.

Diagnostic Methods for Big Survival Data

Yishu Xue

B.A., Central University of Finance and Economics, Beijing, China, 2013

M.A., Columbia University, NY, USA, 2015

A Dissertation
Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy
at the
University of Connecticut

2019

APPROVAL PAGE

Doctor of Philosophy Dissertation

Diagnostic Methods for Big Survival Data

Presented by

Yishu Xue, B.A., M.A.

Co-Major Advisor ______________________________

Dr. Elizabeth D. Schifano

Co-Major Advisor ______________________________

Dr. Jun Yan

Associate Advisor ______________________________

Dr. HaiYing Wang

University of Connecticut

2019

To Nikki

Acknowledgements

First and foremost, I would like to gratefully acknowledge my two advisors, Professor Elizabeth D. Schifano and Professor Jun Yan. I have lost count of how many times you have enlightened, encouraged, guided, and supported me. The time we spent together will be my lifelong treasure.

I would like to thank Professor HaiYing Wang for being my associate advisor in the dissertation committee, serving as my general exam committee member, and our collaboration, which is an important part of this dissertation. I also thank Professor Vladimir Pozdnyakov and Professor Haim Bar for serving as my general exam committee member. My gratitude also goes to Professor Ming-Hui Chen. I benefited a lot from discussions with him in the big data group meetings. Thank you also to Dr. Guanyu Hu for the collaboration and many useful discussions.

I thank all professors who have taught me at UConn. The knowledge I learned from them is invaluable. I would also like to use this special opportunity to thank Professor Arian Maleki at Columbia, whose course inspired my interest in Statistics.

Special thanks to Ms. Tracy Burke, Ms. Megan Petsa, and Mr. Anthony Luis, for their generous administrative assistance.

For the treasure of their friendship, I would like to thank my fellow students in the department. I'm grateful that our paths have crossed. Your support and company made

my PhD journey delightful.

I am also grategul to many of my friends in China for being there for me. Despite that we live in different time zones, your care always comes to me across the ocean.

None of these would be possible without the education and support from my parents, Jinxiu Yan and Yuxiang Xue. I can't thank them enough. I also thank my late grandmother, Chunxiang Xue, for her 16 years of persistent, consistent and unceasing love. I wish she could be here and witness this very important moment. And my dear cat Nikki, thank you for being in my life. Although you know nothing about statistics, your silent but powerful companionship contributed a lot in this dissertation.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 The Cox Model and Its Diagnostics

The Cox model (Cox, 1972) is the most commonly used tool in analyzing survival data and remains so even for massive data (e.g., Mittal et al., 2014). Its usage has extended to fields beyond biostatistics, such as predicting bank failures in finance (Lane et al., 1986), identifying determinants for duration of unemployment in labor market research Kupets (2006), and modelling time until a policy is adopted in political science (Jones and Branton, 2005). It has been deemed one of the "breakthroughs in statistics" (Kotz and Johnson, 1992), and has been cited over 48,848 times up to the time when this dissertation is written.

Due to its pervasive applicability, before taking the results from a fitted Cox model as valid, one should address a few important questions: is the proportional hazards assumption satisfied? Are the functional forms of the variables appropriate? Are there any outliers or influential observations? To answer these questions, multiple methods have been proposed, many of which rely on different types of residuals of the model.

1.2 Cox Model for Big Survival Data

Diagnostics for the Cox model, when raised to the scale of huge datasets, which are not uncommon in this era of information technology, presents challenges to standard statistical analyses. 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). In addition to huge number of observations, there are also examples of survival data from the genomics field that involves gene expression, which usually have a huge number of covariates. In using the Cox model for such datasets, Park and Hastie (2007) proposed a path following algorithm for L_1-regularized generalized linear models that uses a predictor-corrector scheme to find the entire regularization path. They extended this scheme by generalizing the *loss plus penalty* to any convex and differentiable functions, one of which, is the partial likelihood of the Cox model. The **glmpath** package (Park and Hastie, 2018) implements this algorithm. Goeman (2009) proposed a combination of gradient ascent optimization and the Newton–Raphson algorithm that efficiently does L_1-penalized estimation, which can be applied to generalized linear models and the Cox model. Yang and Zou (2013) introduced a mixture of coordinate decent, the majorization-minimization principle and the strong rule to compute the solution paths of the Cox model with elastic net penalty, and implemented it in **fastcox** (Yang and Zou, 2017). Mittal et al. (2014) proposed a variation of coordinate descent that scales for high-dimensional, massive sample-size (HDMSS) data. In their recent work, Wang et al. (2018) proposed a divide-and-conquer algorithm to fit sparse Cox regression on massive-size, moderately-high-dimensional datasets, which greatly improves computational speed and at the same time, maintains similar statistical efficiency as full data based estimators. Nonetheless, little attention has been paid to checking the fundamental assumptions of the Cox model for such huge datasets, which has not been tackled for big data where the data size exceeds a computer's memory.

The rest of this dissertation is organized as follows: Chapter 2 summarizes and reviews diagnostic methods for the Cox model. Different residuals are introduced, and their usage in model diagnostics are discussed, including checking the proportional hazards assumption, verifying functional forms, detecting outlying observations, and identifying influential observations. The diagnostic plots and tests are illustrated with an application regarding dental clinic visits using existing R packages. Chapter 3 presents the construction and asymptotic properties of the online updating cumulative and window version test statistics for the proportional hazards assumption. Under extensive simulation studies, they prove to hold their sizes when proportionality holds, and have substantial power when it is violated in two different ways. The application of this method on lymphoma cancer patients in the Surveillance, Epidemiology, and End Results Program (SEER) is presented. Chapter 4 extends the diagnostics to the nonparametric baseline hazard component of the Cox model. Ideas from statistical process control and statistical profile monitoring are used to design an integrated filtering rule that identifies changes in either part of the Cox model. This dissertation is concluded with a discussion of proposed methods and directions for future research.

Chapter 2

Review on Diagnostics for Cox Model

2.1 Preliminaries

Let T_i^* be the true event time and C_i be the censoring time for subject i such that T_i^* and C_i are independent. Define $T_i = \min(T_i^*, C_i)$ and $\delta_i = I(T_i^* \leq C_i)$, i.e., an indicator that equals 1 if the observation is not censored. Suppose we observe independent copies of (δ_i, T_i, X_i), $i = 1, \ldots, n$, where X_i is the p-dimensional vector of covariates of the ith subject. The Cox model specifies the hazard for individual i as

$$\lambda_i(t) = \lambda_0(t) \exp\left(X_i^\top \beta\right), \tag{2.1}$$

where λ_0 is an unspecified non-negative function of time called the baseline hazard, and β is a p-dimensional coefficient vector in a compact parameter space. Because the hazard ratio for two subjects with fixed covariate vectors X_i and X_j,

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \frac{\lambda_0(t)\exp\left(X_i^\top \beta\right)}{\lambda_0(t)\exp\left(X_j^\top \beta\right)} = \exp\left\{(X_i - X_j)^\top \beta\right\},$$

is a constant over time, and is exponentially proportional to the difference of X_i and X_j, the model is also known as the proportional hazards model. In the case of a single binary predictor, β summarizes the hazard ratio between the corresponding two subgroups of data. It has been later extended to incorporate time-dependent covariates. For the rest of this dissertation, we use $X_i(t)$ to indicate the possibility of covariates being time-dependent.

Cox (1972, 1975) formulated the partial likelihood approach to estimate β. For untied failure time data, Fleming and Harrington (1991) expressed the partial likelihood under the counting process formulation to be

$$\mathrm{PL}(\beta) = \prod_{i=1}^{n} \prod_{t \geq 0} \left[\frac{Y_i(t)\exp\left\{X_j(t)^\top \beta\right\}}{\sum_{j=1}^{n} Y_j(t)\exp\left\{X_j(t)^\top \beta\right\}} \right]^{\mathrm{d}N_i(t)}, \tag{2.2}$$

where $Y_i(t) = I(T_i \geq t)$ 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 $\sum_{i=1}^{n} \mathrm{d}N_i(t) \leq 1$ for any t.

Taking the natural logarithm of (2.2) gives the log partial likelihood in the form of a summation:

$$pl(\beta) = \sum_{i=1}^{n} \int_{0}^{\infty} \left[Y_i(t) \exp\left\{X_i(t)^\top \beta\right\} - \log \sum_{j=1}^{n} Y_j(t) \exp\left\{X_j(t)^\top \beta\right\} \right] \mathrm{d}N_i(t). \tag{2.3}$$

Differentiating (2.3) with respect to β yields the $p \times 1$ score vector $U(\beta)$:

$$U(\beta) = \sum_{i=1}^{n} \int_{0}^{\infty} \left[X_i(t) - \overline{X}(\beta, t)\right] \mathrm{d}N_i(t), \tag{2.4}$$

where $\overline{X}(\beta, t)$ is a weighted mean of covariates for those observations still at risk at time t with the weights being their corresponding risk scores, $\exp\{X_i(t)^\top \beta\}$,

$$\overline{X}(\beta, t) = \frac{\sum_{i=1}^{n} Y_i(t) \exp\left\{X_i(t)^\top \beta\right\} X_i(t)}{\sum_{i=1}^{n} Y_i(t) \exp\left\{X_i(t)^\top \beta\right\}}. \tag{2.5}$$

Taking the negative second order derivative of $pl(\beta)$ yields the $p \times p$ observed partial information matrix

$$\mathcal{I}_n(\beta) = \sum_{i=1}^{n} \int_{0}^{\infty} V(\beta, t) \mathrm{d}N_i(t),$$

with $V(\beta, t)$ being the weighted variance of X at time t:

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

The maximum partial likelihood estimator $\widehat{\beta}_n$ is obtained as the solution of $U(\beta) = 0$.

The solution $\widehat{\beta}_n$ is consistent, and asymptotically normal, i.e., its distribution is approximated by a normal distribution with mean β_0 and inverse variance-covariance matrix being $\mathrm{E}\{\mathcal{I}_n(\beta_0)\}$, where β_0 denotes the true underlying parameter. The evaluation of the expectation depends on extra information which is generally unavailable. The observed information $\mathcal{I}_n(\widehat{\beta}_n)$, however, can be computed using

$$\mathcal{I}_n(\widehat{\beta}_n) = - \left. \frac{\partial^2 pl(\beta)}{\partial\beta\partial\beta^\top} \right|_{\beta=\widehat{\beta}_n}, \tag{2.6}$$

which approximates the variance of $\widehat{\beta}_n$.

Notice that in this section, untied event times are assumed. There are several methods to handle ties, including the Breslow approximation, the Efron approximation, the exact partial likelihood, and the average likelihood methods. In R, the Efron approximation is the default becauseit can be easily implemented, and returns fairly accurate results. For more details, see Section 3.3 of Therneau and Grambsch (2000).

2.2 Proportional Hazards Assumption

2.2.1 Diagnostics Based on Schoenfeld Residuals

Schoenfeld (1980) proposed a chi-squared goodness-of-fit test statistic for the proportional hazards regression model which utilized a residual of the form *Expected - Observed*. The formal definition and its properties were later discussed in Schoenfeld (1982).

Let d denote the total number of events, and we denote the ordered uncensored event times from smallest to largest as $t_1, \ldots, t_d$. Let $X_{(\ell)}$, $\ell = 1, \ldots, d$ be the covariate vector of a subject with an event at the t_ℓ. Further let R_ℓ denote the the risk set at time t_ℓ, which is the set of all individuals who are still alive ("at risk") at t_ℓ. The Schoenfeld residual is defined as

$$r_\ell(\beta) = X_{(\ell)} - \mathrm{E}(X_{(\ell)}|R_\ell), \quad \ell = 1, \ldots, d \tag{2.7}$$

which, when there are no tied event times, is indeed $r_\ell(\beta) = X_{(\ell)} - \overline{X}(\beta, t_\ell)$, where $\overline{X}(\beta, t_\ell)$ as given in Equation (2.5) is evaluated at and t_ℓ. In practice, we replace β with $\widehat{\beta}_n$ and obtain $\widehat{r}_\ell$. If the proportional hazards assumption holds, $E(\widehat{r}_\ell) \doteqdot 0$. Therefore, a plot of Schoenfeld residuals against event times will approximately scatter around 0.

Moreau et al. (1985) and Moreau et al. (1986) proposed a test statistic for goodness-of-fit of the Cox model, with the alternative model being one having time-varying coefficients. In the case of fitting a model with a single covariate in several levels, the statistic is of a sum of quadratic expressions, and reduces to the statistic in Schoenfeld (1980) for two-level problems, but is computationally simpler.

Grambsch and Therneau (1994) generalized the approach in Schoenfeld (1982) to test the proportional hazards assumption. Assuming the true hazard function is of the time-varying form

$$\beta_j(t) \equiv \beta_j + \theta_j g_j(t), \quad j = 1 \ldots, p, \tag{2.8}$$

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, and the natural logarithm transformation. Writing the true hazard function (2.8) using matrix notation, we have

$$\lambda_i(t) = \lambda_0(t) \exp\left[X_i(t)^\top \{\beta + G(t)\theta\}\right], \quad i = 1, \ldots, n, \tag{2.9}$$

where $G(t)$ is a $p \times p$ diagonal matrix with the jth diagonal element being $g_j(t)$, and $\theta = (\theta_1, \ldots, \theta_p)^\top$ is a vector of scalars. Then the null hypothesis of β being time-invariant becomes $H_0 : \theta = 0_{p\times 1}$. Denote $\widehat{V}_\ell = V(\widehat{\beta}_n, t_\ell)$, $G_\ell = G(t_\ell)$, and let

$$H = \sum_{\ell=1}^{d} G_\ell \widehat{V}_\ell G_\ell^\top - \left(\sum_{\ell=1}^{d} G_\ell \widehat{V}_\ell\right)\left(\sum_{\ell=1}^{d} \widehat{V}_\ell\right)^{-1}\left(\sum_{\ell=1}^{d} G_\ell \widehat{V}_\ell\right)^\top.$$

Grambsch and Therneau (1994) proposed the statistic

$$T(G) = \left(\sum_{\ell=1}^{d} G_\ell \widehat{r}_\ell\right)^\top H^{-1} \left(\sum_{\ell=1}^{d} G_\ell \widehat{r}_\ell\right), \tag{2.10}$$

which, under the null hypothesis, has asymptotic distribution χ_p^2, i.e., chi-squared distribution with p degrees of freedom. They also pointed out that the tests in other previous works fall under this framework with different choices of $G(t)$. Table 1 summarizes the related publications and the form of $G(t)$ they used. The form of $G(t)$ is diagonal for all the articles, so we refer to a univariate $g(t)$.

For identifiability, $g(t)$ is assumed to vary around 0, so for data analysis G_ℓ, $\ell = 1, \ldots, d$, need to be centered such that $\sum_{\ell=1}^{d} G_\ell = 0$. In addition, it has been pointed out by Therneau and Grambsch (2000) that $\widehat{V}_\ell$ is rather stable for most datasets, and therefore $\sum_{\ell=1}^{d} G_\ell \widehat{V}_\ell$ is often small. As a result, H is often replaced by

$$H = \sum_{\ell=1}^{d} G_\ell \widehat{V}_\ell G_\ell^\top .$$

The cox.zph() function in the R **survival** package implements the test in (2.10) using this same centering technique. For the rest of this article, we will assume that all G matrices are centered prior to any calculation of diagnostic statistics. User-defined forms of $g(t)$ in obtaining $T(G)$ is also supported. The function also provides a univariate version test for each covariate j as

$$T_j(g) = \left(\sum_{\ell=1}^{d} g_j \widehat{r}_{j\ell} \right)^2 / H_{jj}, \quad j = 1, \ldots, p,$$

where g_j and H_{jj} are the jth diagonal elements of $G(t)$ and H, respectively, and $\widehat{r}_{j\ell}$ is the j th element of $\widehat{r}_\ell$. The test statistic will have a χ_1^2 distribution if the proportional hazards assumption for the jth covariate is satisfied.

Park and Hendry (2015) showed that the decision of time transformations can have profound implications for the conclusions reached. In addition, they suggested that prior to fitting the model, practitioners should first determine the levels of censoring in their

Table 1: Articles and their functional forms of $g(t)$ falling under the framework of Grambsch and Therneau (1994).

Article	$g(t)$
Cox (1972), Gill and Schumacher (1987), Chappell (1992)	a specified function of time
Schoenfeld (1980), Moreau et al. (1985),O'Quigley and Pessione (1989)	piecewise constant on non-overlapping time intervals with the constants and intervals predetermined
Harrell (1986)	$g(t) = \overline{N}(t-)$, tests the correlation between the rank of the event times and the Schoenfeld residuals
Lin (1991)	the proposed test is equivalent to $g(t) = t$ when the maximizer of a weighted partial likelihood, $\widehat{\beta}_w$, is based on a one-step Newton-Raphson algorithm staring from $\widehat{\beta}$
Nagelkerke et al. (1984)	let $g_j(t_1) = 0$ and $g_j(k+1) = a_j^2 \widehat{r}_{jk}$, $j = 1, \ldots, p$ to test for the serial correlation of the Schoenfeld residuals, where a_j is the weight of the jth covariate

data, as in some cases an alternative model might be more appropriate than the Cox model. Exploratory graphical analysis, such as histograms, should be used to see if there are any outlying survival times. If there are few outliers, the test of Grambsch and Therneau (1994) should be done using the untransformed time. Otherwise, the rank transformation is a better choice. They showed using simulations that, with low levels of censoring, the rank and the KM transformation perform approximately equally well. When the level of censoring increases, the rank transformation tends to outperform the KM and natural log transformations.

Keele (2010) pointed out that, while the test of Therneau and Grambsch has been widely used as it is easy to conduct and interpret, application of the test requires some care due to it being sensitive to several forms of misspecification. Omitted predictors, omitted interactions and nonlinear covariate functional forms can all significantly affect the test result. The paper also emphasized the importance of correcting the functional form for continuous covariates before checking for nonproportionality (see Section 2.3).

Winnett and Sasieni (2001) discussed situations in which the approach of Grambsch and Therneau (1994) might provide misleading estimates of time-varying coefficients and presented an example using Mayo clinic lung cancer data. They also suggested using a compromise between $\widehat{V}_\ell$ and $\overline{V}$ for such situations, such as a smoothed version of $\widehat{V}_\ell$. Despite the fact that the test of Grambsch and Therneau (1994) allows for time-dependent covariates, Grant et al. (2014) showed using simulation that its performance, when there are indeed time-dependent covariates, is highly unstable and its power depends largely upon factors that are unknown in practice, such as when the hazard ratio changes, and by how much it changes. Grant et al. (2014) focused on the identity, log, rank, and KM transformations for $g(t)$ in their simulations, and concluded that this instability suggests limited value of the test in (2.10) in the presence of time-dependent covariates in real-world applications. Fisher et al. (1999) suggests the approach of Lin (1991) for time-dependent covariates, but note that the approach can be sensitive to the choice of weight function. Fisher et al. (1999) also cites the approach of Wei (1984), which is based on the score process. Please see Wei (1984) for further details.

Xue et al. (2013) extended the Schoenfeld residuals to case-cohort studies in epidemiological studies of rare disease and defined case-cohort Schoenfeld residuals as the difference of the covariate value and its mean, conditioned on the case-cohort risk set. They also made proper adjustments to the KM estimating procedure by taking into account the influence of each cohort on the increment of the cumulative hazard. They also proposed a test of proportionality based on the correlation between their modified Schoenfeld residuals and $g(t)$, where g could be the identity, rank, or KM transformation. If proportionality holds for a covariate, the correlation should be close to 0. Large values of correlation, however, are often indications of nonproportionality.

2.2.2 Diagnostics Based on Cox–Snell Residuals

Another residual that assists in evaluating the proportional hazards assumption is the Cox–Snell residual. Cox and Snell (1968) provided a general definition of residuals instead of limiting the scope to only linear models. Kay (1977) used the methods in Cox and Snell (1968) to derive the residuals for the proportional hazards regression model. The Cox-Snell residual for the ith observation is defined as:

$$\widehat{e}_i = \widehat{\Lambda}_0(t_i) \exp\left\{X_i(t)^\top \widehat{\beta}_n\right\}, \quad i = 1, \ldots, n, \tag{2.11}$$

where $\widehat{\Lambda}_0$ is the estimated cumulative baseline hazard, which can be obtained using the method of Breslow (1972). More detail on the Breslow estimator is included in Chapter 4. It was concluded that if the model was correctly specified, and no observation was censored, the residuals should approximately exhibit the properties of a random sample of size n from a unit exponential distribution. This can be checked using an exponential Quantile-Quantile plot. Crowley and Hu (1977) used heart transplant survival data to illustrate the usage of Cox–Snell residuals. When censoring is present, however, the residuals are no longer approximately unit exponential.

2.2.3 Diagnostics Based on Martingale Residuals

The martingale residual, which is a slight modification of Cox–Snell residual, also assists in assessing proportionality. It was first discussed by Lagakos (1981) and later by Barlow and Prentice (1988). Further work was done by Therneau et al. (1990). The martingale residual process is defined as

$$\widehat{M_i}(t) = N_i(t) - \int_0^t Y_i(s) \exp\left\{X_i(s)^\top \widehat{\beta}_n\right\} \mathrm{d}\widehat{\Lambda}_0(s), \quad i = 1, \ldots, n, \tag{2.12}$$

where $N_i(t)$ and $Y_i(s)$ are defined in Section 2.1. The martingale residual is defined as the martingale residual process at the end of the study, i.e.,

$$\widehat{M_i} = \delta_i - \int_0^\infty Y_i(s) \exp\left\{X_i(s)^\top \widehat{\beta}_n\right\} \mathrm{d}\widehat{\Lambda}_0(s), \quad i = 1, \ldots, n. \tag{2.13}$$

Asymptotically, $E(\widehat{M_i}) = 0$ and $\mathrm{Cov}(\widehat{M_i}, \widehat{M_j}) = 0$ for $i \neq j$.

Lin et al. (1993) presented a procedure that used cumulative sums of martingale-based residuals, which have been sorted in advance by the order of follow-up time and/or value of a covariate. They considered the process

$$W(z) = \sum_{i=1}^n I(X_i^\top \widehat{\beta}_n < z)\widehat{M_i}, \tag{2.14}$$

which will be an approximate Gaussian process and fluctuate around 0 if the Cox model has been correctly specified. One can perform more formal tests to assess normality (e.g., Kolmogorov-Smirnov, Cramér-von Mises, Anderson-Darling). The authors also discussed the application of such technique to the setting of time-dependent covariates, while arguing that the practical use is little.

Grønnesby and Borgan (1996) concluded that when β is one-dimensional, (2.14) only checks the coding of the covariate. When β is of higher dimension, however, $W(z)$ cannot detect whether the effects of covariates vanish with time or not. Grønnesby and Borgan (1996) grouped the individuals after their linear predictions, i.e., replaced $I(X_i^\top \widehat{\beta}_n < z)$ with $I(X_i^\top \widehat{\beta}_n \in \Omega_\ell)$ in (2.14) for some interval Ω_ℓ, which usually is a quartile group. This is equivalent to introducing the $g \times n$ grouping matrix Q, where g is the number of intervals and $Q_{\ell,i} = I(X_i^\top \widehat{\beta}_n \in \Omega_\ell)$. Given the asymptotic distribution of the estimated martingale residuals, the grouped martingale residual process, $J(\cdot) = Q\widehat{M}(\cdot)$, once properly normalized, converges to a mean zero multivariate Gaussian process. Then with $\widehat{\Sigma}(t)$, such that $\widehat{\Sigma}_{ij}$ is an estimate of the covariance between $J_{\Omega_i}(t)$ and $J_{\Omega_j}(t)$, the test statistic

$$T_C(t) = (J_{\Omega_1}(t), \ldots, J_{\Omega_g}(t))\widehat{\Sigma}^{-1}(t)(J_{\Omega_1}(t), \ldots, J_{\Omega_g}(t))'$$

has an approximate χ^2_{g-1} distribution when the proportional hazards assumption holds.

Marzec and Marzec (1997a) established the asymptotic behavior of processes based on sums of weighted martingale-transformed residuals. They developed Kolmogorov-Smirnov and Cramér-von Mises types of omnibus tests using the fact that, in special cases, they appear to be transformed Brownian motions or Brownian bridges. As the derivation is complicated, please see Marzec and Marzec (1997a) for further details.

2.2.4 Graphical Methods

In addition to formal tests, graphical methods to assess the proportional hazards assumption for categorical predictors have been developed by Cox (1979) and Arjas (1988). Hess (1995) summarized these methods and their extensions, including 1) plotting the Cox model's estimated survival curves $\widehat{S}(t)$ against nonparametric (e.g., Kaplan–Meier) estimates; 2) plotting the estimated cumulative hazard functions $-\log \widehat{S}(t)$ against time and checking if their ratio is constant for any given t; 3) plotting the cumulative hazard functions against each other and checking if the slope is constant; 4) plotting the logarithm of the cumulative hazard functions, $\log\Big(-\log \widehat{S}(t)\Big)$, against time and checking if the curves are approximately parallel; 5) plotting the differences in the log cumulative hazard functions against time and checking if the curve of the differences are approximately constant; and 6) plotting the Schoenfeld residuals against time and checking for changes in patterns of scattering.

The aforementioned graphical methods all have one common limitation: they only apply to categorical predictors that have a few levels. If a predictor has many levels or is continuous, the survival curves and cumulative hazard functions would no longer be informative. Therneau and Grambsch (2000) suggested plotting the cumulative Schoenfeld residuals ordered by event times against event times. If the proportional hazards assumption holds, the cumulative sum should be a random walk starting and ending at 0. These plots, however, can be difficult to read.

2.3 Functional Forms

Martingale residuals, defined in Equation (2.13), play an important role in functional form diagnostics. Barlow and Prentice (1988) provided more detailed discussion and illustrated that plots of such residuals may provide insight to the choice of model form. Therneau et al. (1990) discussed the usage of martingale residuals in investigating the functional form of covariates. To examine a particular covariate, they suggest fitting a proportional hazards model omitting that covariate and computing the martingale residuals $\widehat{M_i}$ as given in Equation (2.13). Then a smoothed plot of $\widehat{M_i}$ versus the omitted covariate often gives approximately the correct functional form of the covariate (e.g., linear, quadratic) to place in the exponent of a Cox model. They also pointed out, however, that this plot does not work well when dealing with large covariate effects, and that it requires the covariate of interest to be uncorrelated with other covariates in the model.

Henderson and Milner (1991) noticed that plots of the martingale residuals against time, although useful, can exhibit systematic patterns which are not *a priori* predictable even when the model fails. They suggested two amendment approaches and gave an example for illustration. One approach was to superimpose the estimated mean when plotting residuals, which enables comparison between the observed patterns and the expected patterns. The other approach was to subtract the conditional expected value from each observed residual and scale it using its standard deviation, which could be consistently estimated from the data according to Barlow and Prentice (1988). Then the standardized residuals, when plotted, should be randomly scattered if the model is appropriate.

Grambsch (1995) proposed two aspects from which the martingale residual plot in Therneau et al. (1990) can be improved. One aspect is to modify the martingale plot for counting process data because of the close relationship between counting process models and Poisson regression. Suppose Z is the variable of interest. If a monotonic relationship between Z and the hazard $\lambda_i(t)$ is expected, a log-linear form is often adequate. The model

$$\lambda_i(t) = \exp\left(\sum_{j=1}^{p-1} \beta_j f_i(X_{ij}) + \alpha Z_i\right) \lambda_0(t), \quad i = 1, \ldots, n,$$

is fitted, and the expected count for the ith individual is

$$\widehat{E}_i = \int_0^{T_i} \exp\left(\sum_{j=1}^{p-1} \widehat{\beta}_{nj} f_j(X_{ij}) + \widehat{\alpha} Z_i\right) \widehat{\lambda}_0(t) dt,$$

where $\widehat{\beta}_{nj}$ denotes the jth entry of $\widehat{\beta}_n$, $\widehat{\lambda}_0(t)$ is the estimated baseline hazard, and $\widehat{\alpha}$ is the estimated parameter for Z. The martingale residual in this case would be $\widehat{M_i} = \delta_i - \widehat{E}_i$, and the generalized linear model (GLM) partial residual is given by

$$\frac{\widehat{M_i}}{\widehat{E}_i} + \widehat{\alpha} Z_i.$$

McCullagh and Nelder (1983) recommended plotting the partial residual against Z as an informal check for the correctness of the guess for functional form.

The other aspect mentioned by Grambsch (1995) comes from the penalized likelihood approach of Hastie and Tibshirani (1993). They assumed that the functional form of covariate X_j is an unknown, smooth function f_j, and proposed the alternative formulation

$$\lambda(t) = \lambda_0(t) \exp\left(\sum_{j=1}^{p} f_j(X_j)\right),$$

which enables estimation of all functional forms at the same time. To avoid overfitting, they maximized the penalized partial likelihood with penalty $\sum_{j=1}^{p} \nu_j \int f_j''(s)^2 ds$, where $\nu_j \geq 0,\ j = 1, \ldots, p$, are smoothing parameters that can be tuned. Both approaches lead to approximately the same solution, but the latter is computationally more complex since the optimization is done within the kernel of the partial likelihood.

2.4 Outlying Observations

A plot of martingale residuals against the linear prediction $X_i(t)^\top \widehat{\beta}_n$ or the risk score $\gamma_i(t) = \exp\left\{X_i(t)^\top \widehat{\beta}_n\right\}$ often helps to identify the observations who have died too soon or lived too long, based on the assumed model. Nevertheless, having a range of $(-\infty, 1]$, the martingale residual is often heavily skewed, and may be misleading. Therneau et al. (1990) used a liver disease data set to demonstrate these scenarios, where the martingale residual plot indicated that some observations died too soon while in actuality they were not outliers at all. They pointed out that it is a favorable practice to transform the residuals to a more normal shaped distribution to help assess the prediction accuracies for individual subjects.

Inspired by the deviance residuals for GLM in McCullagh and Nelder (1983), Therneau et al. (1990) introduced the deviance residual for the Cox model:

$$d_i = \mathrm{sgn}(\widehat{M}_i) \left[-2 \left\{ \widehat{M}_i + \delta_i \log\left(\delta_i - \widehat{M}_i \right) \right\} \right]^{\frac{1}{2}}, \quad i = 1, \ldots, n, \tag{2.15}$$

where δ_i is again the non-censoring indicator for subject i. From the functional form, it is apparent that the deviance residual is essentially a transformation of the martingale residual. Therneau et al. (1990) concluded that with less than 25% of censoring, the deviance residual is approximately normally distributed. With censoring rates greater than 40%, too many points will lie near 0 and make the distribution not normal, but the set of residuals is still symmetrized. Plotting d_i against $X_i(t)^\top \widehat{\beta}_n$ or $\gamma_i(t)$ will help identify potential outliers which have deviance residuals with too large absolute values.

Noticing that deviance residuals do not have a reference distribution and the normal approximation can sometimes be unsatisfactory (Fleming and Harrington, 1991), Nardi and Schemper (1999) proposed two new types of residuals: (i) the log-odds residual $L_i = \log\left[S_i(t_i)/\{1 - S_i(t_i)\}\right]$ and (ii) normal deviate residual $\eta_i = \Phi^{-1}\{S_i(t_i)\}, i = 1, \ldots, n$, where Φ^{-1} is the inverse normal cumulative distribution function. Assuming $S_i(\cdot)$ is known, the sampling distribution for L_i is logistic with $E(L_i) = 0$ and $\mathrm{Var}(L_i) = \frac{\pi^2}{3}$, and standard normal for η_i. In practice, we use the predicted survival for observation i, $\widehat{S}_i(t_i)$, to calculate $\widehat{L}_i$ and $\widehat{\eta}_i$, which converge in probability to L_i and η_i, respectively. Based on simulations, they concluded the performances of these two residuals when identifying outliers are better than that of the deviance residual since they are both unimodal, and the empirical distribution of deviance residual often becomes bimodal because of censoring. They suggested that one can use the quantiles of the normal distribution, ±1.64 and ±1.96, and of the logistic distribution ±2.94 and ±3.66, to help identify potential outliers.

2.5 Influential Observations

The score vector U defined in Equation (2.4) is of great importance in influential diagnostics. Again, using the counting process formulation, the score residual for the ith individual is defined to be

$$r_{Ui}(\widehat{\beta}) = \int_0^\infty [X_i - \overline{X}(\widehat{\beta}_n, s)]\mathrm{d}\widehat{M}_i(s), \quad i = 1, \ldots, n, \tag{2.16}$$

where $\overline{X}(\widehat{\beta}_n, s)$ is the $\overline{X}(\beta, s)$ defined in Equation (2.5) evaluated at $\beta = \widehat{\beta}_n$, and $\widehat{M}_i(s)$ is defined in Equation (2.12).

In studying the influence of one observation, a general practice is to delete that observation, fit the model again, and compare the parameter estimates with those of the model fit on the complete data. Nevertheless, the Cox model is conceptually different from linear or generalized linear models in that it involves both parametric and nonparametric estimation. Therefore, an observation could be influential in terms of more than just regression coefficients. We review measures of both in this section.

2.5.1 Influence on Regression Coefficients

Cain and Lange (1984) presented a method for approximating the influence of individual cases on the Cox model's parameter estimates. Let $\widehat{\beta}_n$ be the value of β that maximizes the partial likelihood (2.2) and $\widehat{\beta}_{n(i)}$ denote the estimate of β when observation i is deleted. They approximated $\widehat{\beta}_n - \widehat{\beta}_{n(i)}$ by assigning to observation i weight w_i. Suppose $w_j = 1$ for any $j \neq i$. Then $\widehat{\beta}_n$ can be regarded as a function of w_i and we have $\widehat{\beta}_n(1) = \widehat{\beta}_n$ and $\widehat{\beta}_n(0) = \widehat{\beta}_{n(i)}$. The first-order Taylor series expansion about $w_i = 1$ gives:

$$\widehat{\beta}_n - \widehat{\beta}_{n(i)} \doteq \frac{\partial \widehat{\beta}_n}{\partial w_i}, \quad i = 1, \ldots, n,$$

where $\partial \widehat{\beta}_n / \partial w_i$ is evaluated at $w_i = 1$. They evaluated the derivative treating the score vector U in Equation (2.4) as a function of $\widehat{\beta}_n$ and w_i, and obtained:

$$\frac{\partial U}{\partial \widehat{\beta}_n} \frac{\partial \widehat{\beta}_n}{\partial w_i} + \frac{\partial U}{\partial w_i} = 0.$$

Notice that $\partial U / \partial \widehat{\beta}_n$ is the negative observed information matrix defined in Equation (2.6).

Hence we obtain

$$\frac{\partial \widehat{\beta}_n}{\partial w_i} = \mathcal{I}_n^{-1}(\widehat{\beta}_n)\frac{\partial U}{\partial w_i}. \tag{2.17}$$

The partial derivative $\partial U/\partial w_i$, when evaluated at $w_i = 1$, becomes exactly the score residual r_{Ui} in Equation (2.16). Therefore

$$\left(\frac{\partial \widehat{\beta}_n}{\partial w_i}\right)_{w_i=1} = \mathcal{I}_n^{-1}(\widehat{\beta}_n) r_{Ui}, \quad i = 1, \ldots, n.$$

Let D be the $n \times p$ matrix with the ith row being $\widehat{\beta}_n - \widehat{\beta}_{n(i)}$, and r_U be the $n \times p$ matrix with the ith row being the vector of score residuals for observation i. Then the above approximation, put into matrix form, becomes

$$D = r_U \mathcal{I}_n^{-1}(\widehat{\beta}_n). \tag{2.18}$$

We call D the matrix of *dfbeta* residuals. When we divide D_{ij} by the observed standard deviation of the ith element of $\widehat{\beta}_n$, which is the square root of the ith diagonal element of the inverse observed information matrix $\mathcal{I}_n^{-1}(\widehat{\beta}_n)$, we get D_S, the matrix of *dfbetas* residuals. Conventionally, the ith observation is considered to be influential if $D_{Sij} > 1$ for small to medium datasets, and if $D_{Sij} > 2/\sqrt{n}$ for large datasets.

Reid and Crépeau (1985) presented influence functions for the Cox model to identify possible influential observations and gave the same statistic (2.18) as in Cain and Lange (1984).

Storer and Crowley (1985) pointed out that a good estimate of $\widehat{\beta}_n - \widehat{\beta}_{n(i)}$ can also be obtained using an augmented regression model. The design matrix is augmented using a binary indicator variable for the ith observation and taking a single Newton-Raphson step towards the fit of the augmented model gives the estimate of change in β. This estimate, they argued, is easy to compute.

2.5.2 Overall Influence

Pettitt and Daud (1989) discussed the disadvantages of the approaches that try to approximate $\widehat{\beta}_n - \widehat{\beta}_{n(i)}$. They concluded that only using single-case deletion statistics may cause some cases to be masked, i.e., the deleted observation may influence the value of the test statistic enough so that an actual outlier is not declared as outlier. They suggested changing the weights of each observation, and studying the change in the likelihood caused by this perturbation. They adopted the approach of Cook (1986) and defined the likelihood displacement to be

$$\mathrm{LD}(w) = 2\left[\ell\left(\widehat{\beta}_n\right) - \ell\left(\widehat{\beta}_n(w)\right)\right], \tag{2.19}$$

where $\widehat{\beta}_n(w)$ maximizes the weighted partial likelihood

$$\mathrm{PL}_w(\beta) = \prod_{i=1}^{n}\prod_{t\geq 0}\left[\frac{w_i Y_i(t)\exp\left\{X_i(t)^\top\beta\right\}}{\sum_{j=1}^{n} w_j Y_j(t)\exp\left\{X_j(t)^\top\beta\right\}}\right]^{\mathrm{d}N_i}. \tag{2.20}$$

The weighting scheme of $w_i = 1, i \neq j$ and $w_j = 0$ in Cain and Lange (1984) is an appropriate and specific case. Using second-order approximation, we have

$$\ell(\widehat{\beta}_n) - \ell(\widehat{\beta}_n(w)) \approx \frac{1}{2}\left[\widehat{\beta}_n - \widehat{\beta}_n(w)\right]^\top \mathcal{I}_n(\widehat{\beta}_n)\left[\widehat{\beta}_n - \widehat{\beta}_n(w)\right].$$

Let $U_w(\beta)$ be the score function corresponding to the weighted partial log-likelihood. With another approximation that

$$\frac{\partial \widehat{\beta}_n(w)}{\partial w^\top} = \mathcal{I}_n^{-1}(\widehat{\beta}_n)\frac{\partial U_w(\beta)}{\partial w^\top},$$

which is essentially the matrix form of Equation (2.17), LD(w) reduces to

$$\mathrm{LD}(w) \approx (w_0 - w)^\top r_U \mathcal{I}_n^{-1}(\widehat{\beta}_n) r_U^\top (w_0 - w), \tag{2.21}$$

where w_0 is a vector of 1's and r_U is the score residual matrix. The approach of Cook (1986) looks for an unit-length $l_{n\times 1}$ that maximizes $l^\top B l$, where $B = r_U \mathcal{I}_n^{-1}(\widehat{\beta}_n) r_U^\top$. The maximum $\xi_{\max}$ is the largest eigenvalue of B, and is attained when $l_{\max}$ is the corresponding eigenvector. Cook concluded that $\xi_{\max} > 1$ indicates notable local sensitivity, and that a locally influential observation must be globally influential, although the reverse is not necessarily true.

Weissfeld (1990) adopted the idea of Cook (1986) to measure the change in likelihood function by computing its curvature. Originally, in Cook's work, the change could be caused by perturbations in the score vector or the covariates. Weissfeld (1990) proposed for the Cox model three ways to perturb the data: weighting the observations in the log partial likelihood using a vector w of weights, adding a vector to the vector of censoring indicators $(\delta_1, \ldots, \delta_n)$, and adding a scaled weight vector w to the covariates, where the scale is usually the standard deviation of the corresponding coefficient. Then take $\ddot{F} = \Delta^\top \mathcal{I}_n^{-1}(\widehat{\beta}_n)\Delta$, where $\mathcal{I}_n^{-1}(\widehat{\beta}_n)$ is the inverse of the observed information matrix in Equation (2.6) and Δ is the partial derivative matrix of the score vector U to the weights, which takes different forms for the three perturbation schemes. The maximum eigenvalue of $\ddot{F}$, $C_{\max}$, is informative in that large or small values point to possible influential observations. It was concluded that perturbation of the covariates is useful for locating observations that influence the estimated coefficients, and the other two pertubations will help detect observations that may impact the results of likelihood ratio tests. It was also indicated that the proposed approach is capable of detecting influential observations caused by masking.

Barlow (1997) proposed a modification of the method in Pettitt and Daud (1989). Their approach replaces $\mathcal{I}_n(\widehat{\beta}_n)$ in Equation (2.21) using the inverse of the robust covariance matrix in Lin and Wei (1989). The substitution, upon further derivation, provides a scalar measure of influence with known mean to be the ratio of number of events and number of observations, and range of (0,1). The approach can also be generalized to include designs with multiple failures and to case-cohort designs. They illustrated the usage of this method by plotting the calculated influence measure against the covariate of interest, and visually looking for any particularly influential observations.

In addition to traditional delete-one approaches, Wei and Kosorok (2000) developed case interaction influence measures for unmasking observations masked by other observations in the Cox model. They proposed the following statistic to assess the joint influence of observations i and j:

$$\begin{aligned} -\left(\widehat{\beta}_n - \widehat{\beta}_{n(j)} - \widehat{\beta}_{n(i)} + \widehat{\beta}_{n(i,j)}\right) &= \left(\widehat{\beta}_{n(i)} - \widehat{\beta}_{n(i,j)}\right) - \left(\widehat{\beta}_n - \widehat{\beta}_{n(j)}\right) \\ &= \left(\widehat{\beta}_{n(j)} - \widehat{\beta}_{n(i,j)}\right) - \left(\widehat{\beta}_n - \widehat{\beta}_{n(i)}\right) \\ &= \left(\widehat{\beta}_n - \widehat{\beta}_{n(i,j)}\right) - \left\{\left(\widehat{\beta}_n - \widehat{\beta}_{n(i)}\right) + \left(\widehat{\beta}_n - \widehat{\beta}_{n(j)}\right)\right\}, \end{aligned}$$

where $\widehat{\beta}_n - \widehat{\beta}_{n(i,j)}$ and $\widehat{\beta}_{n(i)} - \widehat{\beta}_{n(i,j)}$ are related to the joint influence and conditional influence in Lawrance (1995). On one hand, if the value of the test statistic is small, we conclude that the parameter estimate is not significantly influenced by the deletion of one observation, with or without incorporating the other observation in estimation. A large value, on the other hand, would imply that the joint influence of these two observations is significantly different from the sum of their individual influences, and the identified pairs need further investigation. In cases where two moderately influential observations have substantial joint influence, or where two individually influential observations have little joint influence, however, their proposed diagnostic cannot identify them.

Zhu et al. (2015) investigated case-deletion measures, conditional martingale residuals, and score residuals for the Cox model with missing covariate values. They proposed the Q-distance to examine the effects of deleting individual observations on the estimates of finite-dimensional and infinite-dimensional parameters. They also addressed the problem of quantifying influence by introducing a detection probability of being influential for each observation and for any case-deletion measure. A large value of detection probability is an indicator of being influential. The forms and derivation of the Q-distance and the detection probability are complicated; the interested reader should see Zhu et al. (2015) for full details.

2.6 A Case Study

The dental restoration longevity data, provided by the University of Iowa College of Dentistry's Geriatric and Special Needs (SPEC) Clinic (see Caplan et al., 2019) is used as a case study to demonstrate the diagnostic methods of the Cox model. For this analysis, electronic data was obtained during the 5-year period from 1995 to 1999. The health record numbers were scrambled by IT personnel to ensure that no Personal Health Information was included. Subsequently, the Institutional Review Board at the University of Iowa declared that this project is exempt from Human Subjects Review, due to the anonymous nature of the data.

We identified 697 unique patients who went to the SPEC Clinic to treat their molars upon their first visit and received restoration in amalgam, composite, or glass ionomer. The follow-up of their visits began on the date of restoration. Any restoration that was replaced with another intracoronal or extracoronal restoration, accessed for endodontic therapy, or extracted was deemed to have undergone an event. If the restoration results in an event, the event date would become the end of follow-up. Restorations that did not incur an event are considered censored up to the date of the patient's second-to-last visit to any College of Dentistry's clinic. Among the 697 patients, 228 experienced an event during the follow-up, giving a censoring rate of 67.3%.

We considered the following covariates: Gender, Age (when receiving restoration, centered and scaled), Occupation (Faculty, Non-faculty) and Size (Small, Medium, Large). Analysis was performed using the **survival** package in R, and figures were produced using the **survminer** (Kassambara and Kosinski, 2017), **ggplot2** (Wickham, 2009) and **ggfortify** (Tang et al., 2016) packages.

2.6.1 Functional Form

As suggested in Section 2.3, we should determine the appropriate form of covariates to include in the model before testing for proportionality. Age is the only continuous covariate whose form needs to be assessed. We use the methods of Therneau et al. (1990): fit a model excluding Age and obtain its martingale residuals. The martingale residuals are plotted against Age in Figure 1. We also superimpose the loess pointwise confidence band. The curvy behavior of the loess fit indicates that we should consider higher orders of Age.

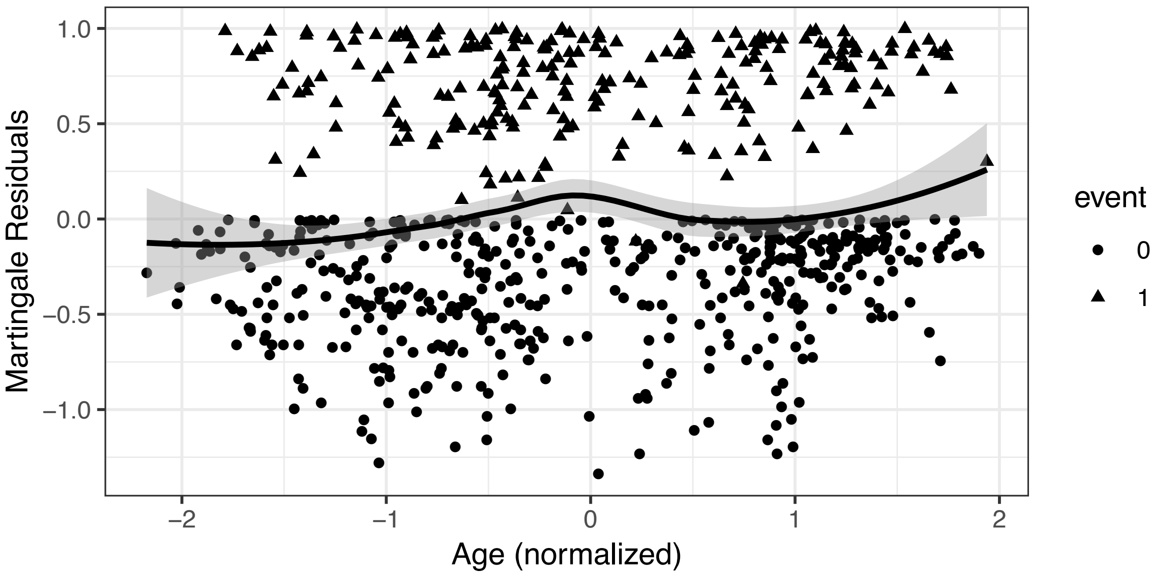

Figure 1: Plot of martingale residual of the model excluding Age against Age.

2.6.2 Proportional Hazards

As suggested in Section 2.6.1, we consider including the square of Age (Age^2) in the model. We fit two models: one with only linear age effects (Model 1) and another model with linear and quadratic age effects (Model 2) to assess the improvement to the model when correcting the functional form. To assess the proportional hazards assumption, we used the cox.zph() function to obtain both the individual χ_1^2 statistics for each covariate and the global χ_p^2 statistic for each model. The test results are summarized in Table 2. While both models passed the global test, Age in Model 1 did not pass the individual test at the 0.05 level. In Model 2, however, both Age and Age^2 pass the individual test of proportionality at the 0.05 level.

The parameter estimates of Model 2 are summarized in Table 3. Restorations for

Table 2: Proportionality test results for Model 1 and Model 2.

	Model 1		Model 2	
	χ^2 Stat	p-value	χ^2 Stat	p-value
Male	0.586	0.444	0.429	0.513
Age	4.029	0.045	3.555	0.059
Age^2	–	–	0.638	0.424
Non-Faculty	0.429	0.513	0.558	0.455
SizeMedium	1.560	0.212	1.711	0.191
SizeSmall	0.298	0.585	0.416	0.519
GLOBAL	6.788	0.237	6.932	0.327

Table 3: Cox regression results for tooth restoration failure for the modified model.

	Estimate	exp(Estimate)	Std.Error	Z Stat	p-value
Male	-0.221	0.802	0.137	-1.612	0.107
Age	0.206	1.228	0.0076	2.709	0.007
Age^2	-0.092	0.912	0.085	-1.075	0.282
Non-Faculty	0.116	1.123	0.146	0.795	0.427
SizeMedium	-0.140	0.869	0.165	-0.850	0.395
SizeSmall	-0.510	0.601	0.169	-3.018	0.003

males tend to fail later than for females, while restorations for older patients tend to fail sooner. Compared to large restorations, medium and small restorations are less likely to fail.

As mentioned in Section 2.2.1, when the proportional hazards assumption holds, the Schoenfeld residuals will be close to zero. Therefore a plot of the Schoenfeld residuals against survival times would be informative (Schoenfeld, 1982). Figure 2 shows the plots for Model 2. For all six covariates, the smoothed pointwise confidence bands are all around 0, which again confirms that there is no obvious evidence against the proportional hazards assumption.

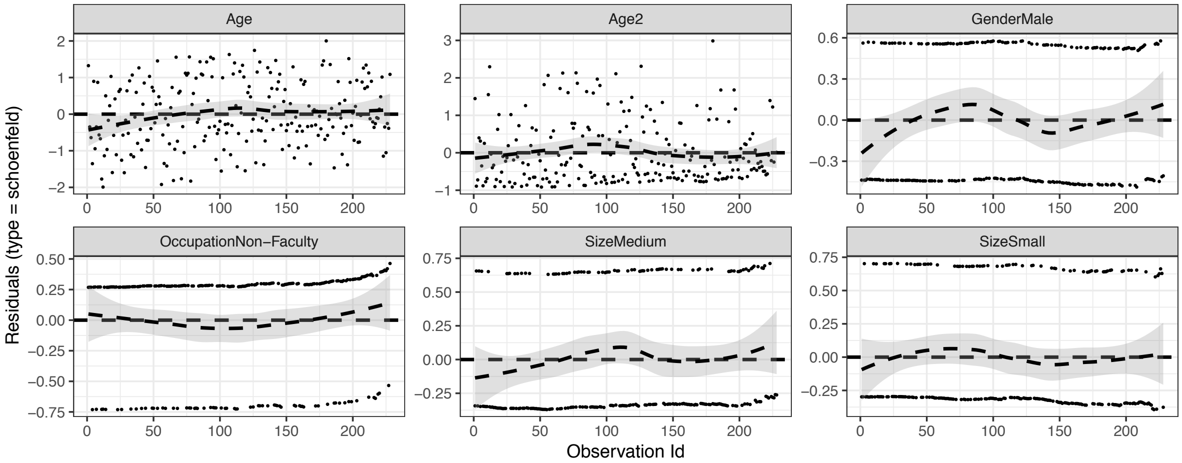

Figure 2: Schoenfeld residuals for each covariate against survival time.

For the three categorical covariates (Gender, Occupation and Size), we also utilize the graphical methods in Section 2.2.4 to check the proportional hazards assumption. For each covariate, we plot the estimated survival curves $\widehat{S}(t)$, the cumulative hazards $-\log\widehat{S}(t)$ and the log-log transformed survival $\log\Big(-\log\widehat{S}(t)\Big)$ against survival times in Figure 3. The three plots for Gender indicate that the hazards of the two gender strata are proportional, but the lack of large discrepancy indicates that this proportional effect is not significant. Similarly, the ignorable discrepancy between the two occupation strata tells the same story. The three plots for restoration size strata, however, are more informative, in that although the proportionality effect is small between SizeLarge and SizeMedium, it is highly significant between SizeLarge and SizeSmall.

As mentioned in Section 2.2, the martingale residual can be used to graphically assess the proportional hazards assumption as well. We plot the cumulative sum of martingale residuals ordered by Age in Figure 4. The curve fluctuates around zero as expected.

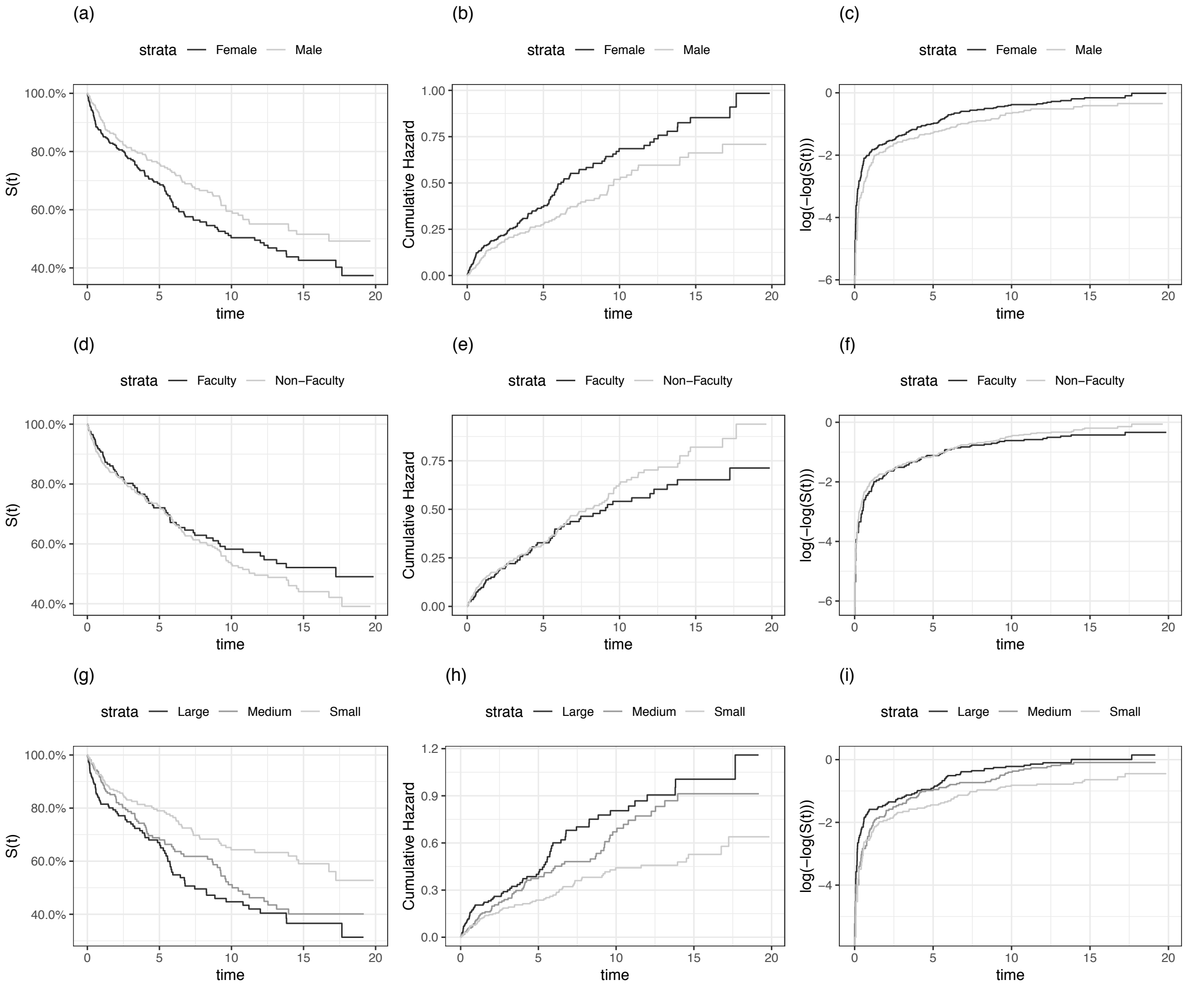

Figure 3: Estimated survival curves, cumulative hazards and log-log transformed survival curves for categorical covariates. The first row is for Gender, the second row is for Occupation, and the third row is for Size.

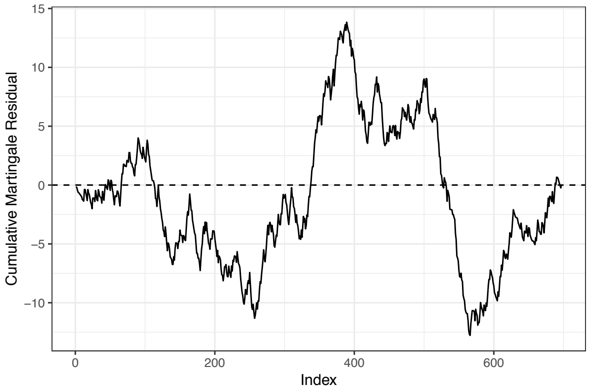

Figure 4: Cumulative sum of martingale residuals of Model 2, ordered by Age.

2.6.3 Outlying Observations

As suggested in Section 2.4, both the martingale residual and the deviance residual are useful for identifying outlying observations, but the deviance residual is less skewed and therefore more useful. We plot both residuals against the linear predictions, $X^\top \widehat{\beta}_n$, in Figure 5. In Figure 5(a), the martingale residuals do not vary much against the linear predictions, and fail to identify any outlying observations. Using ±1.96 as thresholds, the deviance residuals plotted in Figure 5(b) identify 34 potential outliers. Upon further investigation, these subjects turned out to be much younger than other subjects (46.6 vs 55.1) but their restorations failed very soon. Due to the high censoring rate, however, the normal-approximation-based thresholds may not be appropriate.

We also use the log-odds residual and the normal deviate residual discussed in Section 2.4 to look for potential outliers. Both the log-odds residual in Figure 6(a) and the normal deviate residual in Figure 6(b) identify the same set of 67 potential outliers, which is bigger than the set of outliers identified by the deviance residual. This set,

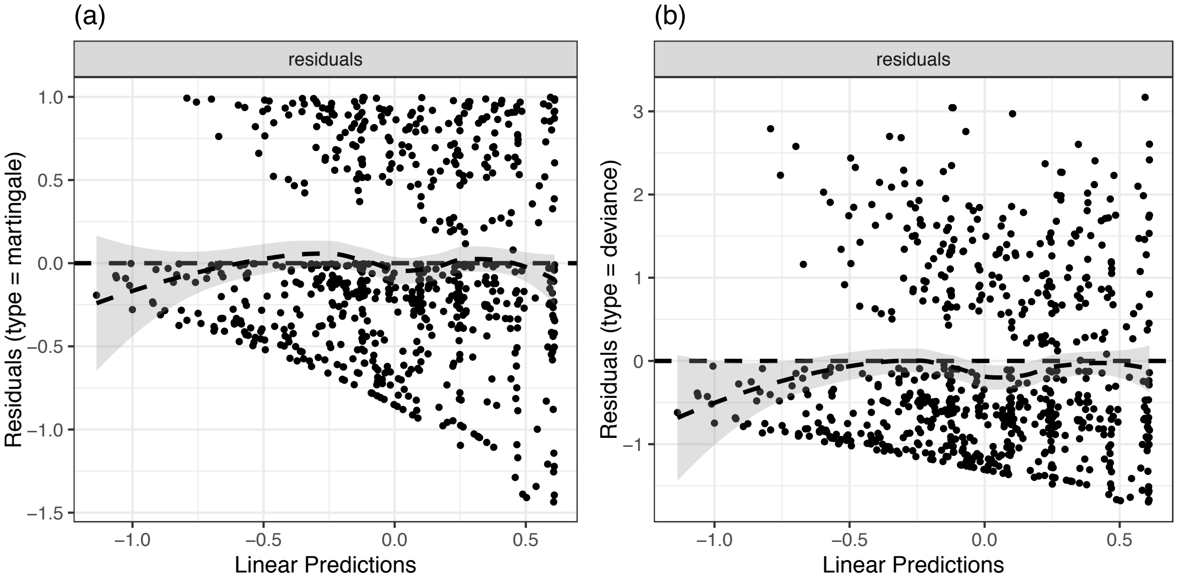

Figure 5: Plot of martingale and deviance residuals of Model 2.

however, still consists of younger individuals (51.7 vs 55.1) whose restorations failed very soon.

2.6.4 Influential Observations

We use the methods in Section 2.5 to perform influential diagnostics. We first look at influence of observations on parameter estimates and plot the *dfbetas* residuals in Figure 7. As illustrated, no observation caused any parameter change of more than 15% of that parameter's standard error. Considering that there are 697 observations, we can conclude there are no significantly influential observations.

We also present the likelihood displacement approach in Figure 8. The absence of particularly large likelihood displacements further confirms our conclusion from Figure 7.

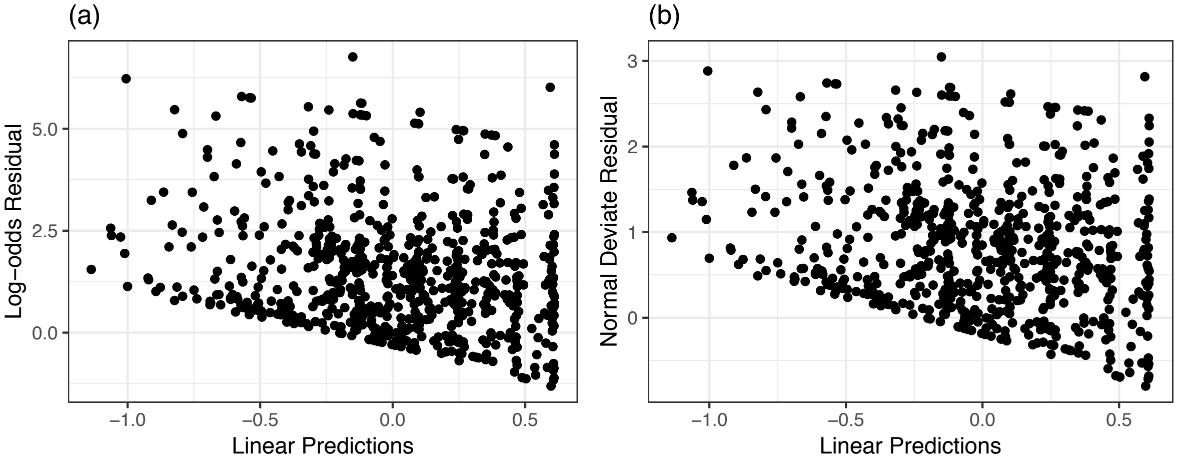

Figure 6: Plot of log-odds and normal deviate residuals of Model 2.

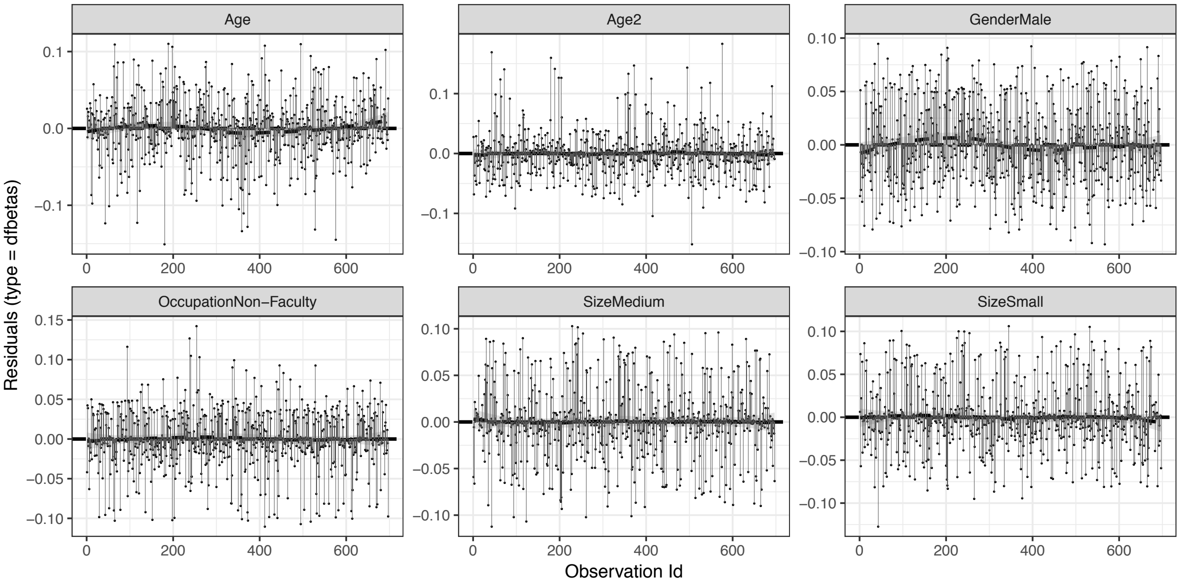

Figure 7: Dfbetas residuals for covariates of Model 2.

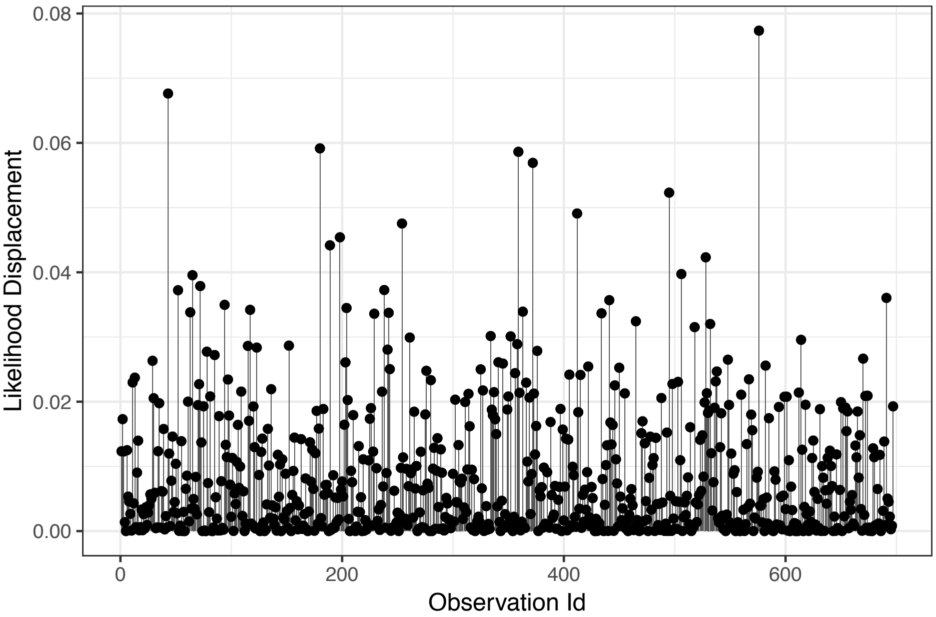

Figure 8: Likelihood displacement caused by dropping each observation.

2.7 Discussion

With such wide usage across a variety of disciplines, the importance of the Cox regression for modeling time-to-event data cannot be overstated. As a consequence, one must consider the appropriateness and validity of the results from such an analysis before reaching at any conclusions. This chapter summarizes existing graphical and statistical diagnostic methods for the Cox model on given, full datasets, including methods for identifying violations of the proportional hazards assumption, finding appropriate functional forms of continuous covariates, and detecting outlying and influential observations. Using a non-linear functional form of covariate can often improve model fit, while any outlying or influential observations identified by the procedures should be investigated further before taking any action.

Violations of the proportional hazards assumption can be addressed in several ways, the most common of which include the use of time-varying coefficients and stratified models. Flexible models that incorporate time-varying coefficients have been studied by Murphy and Sen (1991), Hastie and Tibshirani (1993), Verweij and van Houwelingen (1995), Sargent (1997), Marzec and Marzec (1997b), Cai and Sun (2003), Tian et al. (2005), Fan et al. (2006), and more recently by Chen et al. (2012). In practice, the graphical tools in the **survival** package enable us to check if there are any time-varying coefficients, and the survSplit() function, which will be introduced in more detail in Chapter 3, facilitates the approximation for such coefficients using piecewise constant functions, and we are able to test for excessive time-variation. Another popular approach for addressing non-proportionality is using a stratified Cox model. In this case, it is assumed that individuals in different strata have different baseline hazard functions, but all other predictor variables satisfy the proportional hazards assumption within each stratum. Related chapters can be found in Therneau and Grambsch (2000), Kalbfleisch and Prentice (2002), Lawless (2003), and Collett (2015).

The Cox model has also been extended to the analysis of interval-censored survival data. Such models have been studied by Finkelstein (1986), Farrington (2000), Goggins and Finkelstein (2004) and recently Heller (2010). In particular, Farrington (2000) proposed the counterparts to the Cox–Snell, martingale, deviance, and Schoenfeld residuals and illustrated their usage in model diagnostics under the interval-censored framework.

Chapter 3

Online Updating Proportional Hazards Test

3.1 Introduction

Proportional hazards is the fundamental assumption made by the Cox model. If it is itself violated, neither the parameter estimates nor the inference based on them are trustworthy. As mentioned in Section 2.2.1, the test of Grambsch and Therneau (1994) has been popular since it has been proposed, as it incorporates many existing tests, and provides the flexibility of choosing a survival time transformation when calculating the final χ^2 test statistic. It is, however, worth noticing that we need to compute the Schoenfeld residuals for all observations at once in order to obtain the final statistic, which is impossible when the data size is bigger than the computer's memory and an estimate for β cannot even be obtained.

The same issue also exists in linear regression or generalized linear regression problems. In addition to subsample procedures, which inevitably incur information loss, and divide-and-combine procedures, which require powerful computing resource, Schifano et al. (2016) proposed the online updating procedure, which treats the data as a stream, and process the stream in a blockwise fashion. After a block is processed, only a few summarize statistics are retained, and the data itself can be removed from the memory, freeing up space for the next block. In this way, a huge dataset can be processed using a common computer.

In this chapter, Section 3.2.1 proposes the online updating cumulative version test statistic for the proportional hazards assumption with streams of big survival data. As implied by its name, it utilizes information from all historical data. Section 3.2.2 presents an online updating window version variant of the test that focuses on local changes, using information from most recent blocks. At which estimate of β to evaluate the matrices and residuals in calculating the statistics is addressed in Section 3.2.3. Section 3.2.4 provides theoretical justification for the proposed test statistics. Section 3.3 contains the numerical simulation results for both versions of test statistic under a scenario where the proportional hazards assumption for stream data is satisfied, and two scenarios where it is violated. The savings in computing time and memory usage are also studied. Section 3.4 presents results from survival analysis of lymphoma patients from the Surveillance, Epidemiology, and End Results Program (SEER).

3.2 Online Updating

3.2.1 Cumulative Version

Instead of a given, complete dataset, we now consider a scenario in which data become available in blocks. Suppose that for each newly arriving block k, we observe for n_k subjects, an n_k-dimensional vector of response times, event indicators, and an $n_k \times p$ matrix of covariates, respectively, for $k = 1, \ldots, K$ where K is some terminal accumulation point of interest. Further, denote the number of events in the kth block as d_k. With a given $g(t)$ as in (2.8), we obtain d_k centered $p \times p$ diagonal matrices $G(t_1), \ldots, G(t_{d_k})$ such that $\sum_{\ell=1}^{d_k} G(t_\ell) = 0$. Let $G_{\ell k}$ and $\widehat{r}_{\ell k}$, $\ell = 1, \ldots, d_k$, be the kth block counterpart of previously defined G_ℓ and Schoenfeld residual $\widehat{r}_\ell$, respectively. Without loss of generality, we assume that there is at least one event in each block, and each block-wise observed information matrix $\mathcal{I}_{n_k,k}$ is invertible. Let $V_{\ell k}$ be the variance-covariance matrix of the covariate matrix at the ℓth event time in the kth block. With the approximation that $\widehat{V}_{\ell k} = \mathcal{I}_{n_k,k}/d_k$, where $\mathcal{I}_{n_k,k}$ is evaluated at some estimate of β, we have $\sum_{\ell=1}^{d_k} G_{\ell k}\widehat{V}_{\ell k} = 0$. We will discuss the choice of estimate for β that will be used to evaluate $\mathcal{I}_{n_k,k}$, and also $\widehat{r}_{\ell k}$, in Section 3.2.3.

We denote $H_{d_k,k} = \frac{1}{d_k}\sum_{\ell=1}^{d_k} G_{\ell k}\mathcal{I}_{n_k,k}G_{\ell k}$, and $Q_{d_k,k} = \sum_{\ell=1}^{d_k} G_{\ell k}\widehat{r}_{\ell k}$. Let $H_0 = 0_{p\times p}$, $H_{k-1} = \sum_{i=1}^{k-1} H_{d_i,i}$, $Q_0 = 0_{p\times 1}$, and $Q_{k-1} = \sum_{i=1}^{k-1} Q_{d_i,i}$. Then we have the online updating test statistic given by

$$T_k(G) = Q_k^\top H_k^{-1} Q_k = (Q_{k-1} + Q_{d_k,k})^\top (H_{k-1} + H_{d_k,k})^{-1} (Q_{k-1} + Q_{d_k,k}). \tag{3.1}$$

At each accumulation point k, we need to store H_{k-1} and Q_{k-1} from previous calculations, and compute $H_{d_k,k}$ and $Q_{d_k,k}$ for the current block.

3.2.2 Window Version

The cumulative test statistic takes all historical blocks into consideration, one potential problem of which is that discrepancies from the proportional hazards 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 $w (\geq 1)$, which is tunable, and use summary statistics for all blocks in this window to construct the corresponding test statistic. With $H_{d_k,k}$ and $Q_{d_k,k}$ defined above, we again assume there is at least one event in each block of data. Denoting $H_{k-1}^w = \sum_{i=k-w}^{k-1} H_{d_i,i}$, and $Q_{k-1}^w = \sum_{i=k-w}^{k-1} Q_{d_i,i}$, the window version online updating test statistic for nonproportionality based on the most recent w blocks is:

$$T_k^w(G) = (Q_k^w)^\top (H_k^w)^{-1} Q_k^w. \tag{3.2}$$

In implementation, we only need to store $H_{d_k,k}$ and $Q_{d_k,k}$ 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 Q_k, one $p \times 1$ vector for an estimate of β, one $p \times p$ matrix H_k, and one $p \times p$ estimated covariance matrix of β, 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 not to be large, which also makes the storage of these quantities affordable.

3.2.3 Where to Evaluate the Matrices and Residuals

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

Suppose now we have K subsets of data. The score function for subset k is

$$U_{n_k,k}(\beta) = \sum_{i=1}^{n_k} \int_0^\infty \left[X_i(t) - \overline{X}(\beta, t)\right] \mathrm{d}N_i(t).$$

Denote the solution to $U_{n_k,k}(\beta) = 0$ as $\widehat{\beta}_{n_k,k}$. If we define

$$\mathcal{I}_{n_k,k}(\beta) = -\sum_{i=1}^{n_k} \frac{\partial \int_0^\infty \left[X_i(t) - \overline{X}(\beta, t)\right] \mathrm{d}N_i(t)}{\partial \beta},$$

a Taylor expansion of $-U_{n_k,k}(\beta)$ at $\widehat{\beta}_{n_k,k}$ is given by

$$-U_{n_k,k}(\beta) = \mathcal{I}_{n_k,k}(\widehat{\beta}_{n_k,k})(\beta - \widehat{\beta}_{n_k,k}) + R_{n_k,k}$$

as $U_{n_k,k}(\widehat{\beta}_{n_k,k}) = 0$ and $R_{n_k,k}$ is the remainder term. For notational simplicity, we denote $\mathcal{I}_{n_k,k}(\widehat{\beta}_{n_k,k})$ as $\widehat{\mathcal{I}}_{n_k,k}$ for the rest of this thesis. Without loss of generality, we assume that there is at least one event in each block, and each $\widehat{\mathcal{I}}_{n_k,k}$ is invertible.

Similar to the aggregated estimating equation (AEE) estimator of Lin and Xi (2011) which uses a weighted combination of the subset estimators, the AEE estimator under the Cox model framework is:

$$\widehat{\beta}_N = \left\{\sum_{k=1}^{K} \widehat{\mathcal{I}}_{n_k,k}\right\}^{-1} \sum_{k=1}^{K} \left(\widehat{\mathcal{I}}_{n_k,k}\widehat{\beta}_{n_k,k}\right), \tag{3.3}$$

which is the solution to $\sum_{k=1}^{K} \widehat{\mathcal{I}}_{n_k,k}(\beta - \widehat{\beta}_{n_k,k}) = 0$, with N being the total number of observations at the final accumulation point K. Schifano et al. (2016) provided the variance estimate for the original AEE estimator of Lin and Xi (2011), and under the

Cox model framework it simplifies to

$$\widehat{A}_N = \left\{ \sum_{k=1}^{K} \widehat{\mathcal{I}}_{n_k,k} \right\}^{-1}. \tag{3.4}$$

Following Schifano et al. (2016), the cumulative estimating equation (CEE) estimator for β at accumulation point k under the Cox model framework is

$$\widehat{\beta}_k = \left\{ \widehat{\mathcal{I}}_{k-1} + \widehat{\mathcal{I}}_{n_k,k} \right\}^{-1} \left\{ \widehat{\mathcal{I}}_{k-1}\widehat{\beta}_{k-1} + \widehat{\mathcal{I}}_{n_k,k}\widehat{\beta}_{n_k,k} \right\} \tag{3.5}$$

for $k = 1, 2, \ldots$, where $\widehat{\beta}_0 = 0_{p\times 1}$, $\widehat{\mathcal{I}}_0 = 0_{p\times p}$, and $\widehat{\mathcal{I}}_k = \sum_{i=1}^{k} \widehat{\mathcal{I}}_{n_i,i} = \widehat{\mathcal{I}}_{k-1} + \widehat{\mathcal{I}}_{n_k,k}$. The variance estimator at the kth update simplifies to

$$\widehat{A}_k = \left\{ \mathcal{I}_{k-1} + \mathcal{I}_{n_k,k}(\widehat{\beta}_{n_k,k}) \right\}^{-1}. \tag{3.6}$$

Note that for terminal $k = K$, Equations (3.5) and (3.6) coincide with Equations (3.3) and (3.4), respectively (i.e., AEE=CEE).

As pointed out by Schifano et al. (2016), the CEE estimators are not identical to the estimating equation (EE) estimators (based on the entire sample) in finite sample sizes. Similar to Schifano et al. (2016), we propose a CUEE estimator under the EE framework to better approximate the EE estimators with less bias. Take the Taylor expansion of

$-U_{n_k,k}(\beta)$ around $\breve{\beta}_{n_k,k}$, which will be defined later. We have

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

where $\breve{U}_{n_k,k} = U(\breve{\beta}_{n_k,k})$, $\breve{\mathcal{I}}_{n_k,k} = \mathcal{I}_{n_k,k}(\breve{\beta}_{n_k,k})$, and $\breve{R}_{n_k,k}$ is the remainder term. We now ignore the remainder terms, and sum the first order expansions for blocks $1, \ldots, K$, and set it equal to $0_{p\times 1}$:

$$\sum_{k=1}^{K} -\breve{U}_{n_k,k} + \sum_{k=1}^{K} \breve{\mathcal{I}}_{n_k,k}(\beta - \breve{\beta}_{n_k,k}) = 0_{p\times 1}. \tag{3.7}$$

Then we have the solution to (3.7):

$$\widetilde{\beta}_K = \left\{\sum_{k=1}^{K} \breve{\mathcal{I}}_{n_k,k}\right\}^{-1} \left\{\sum_{k=1}^{K} \breve{\mathcal{I}}_{n_k,k}\breve{\beta}_{n_k,k} + \sum_{k=1}^{K} \breve{U}_{n_k,k}\right\}.$$

The choice of $\breve{\beta}_{n_k,k}$ is subjective. At accumulation point k, it is possible to utilize information at the previous accumulation point $k-1$ to define $\breve{\beta}_{n_k,k}$. One candidate intermediary estimator can be obtained as

$$\breve{\beta}_{n_k,k} = (\breve{\mathcal{I}}_{k-1} + \widehat{\mathcal{I}}_{n_k,k})^{-1} \left(\sum_{i=1}^{k-1} \breve{\mathcal{I}}_{n_i,i}\breve{\beta}_{n_i,i} + \widehat{\mathcal{I}}_{n_k,k}\widehat{\beta}_{n_k,k}\right) \tag{3.8}$$

for $k = 1, 2, \ldots$, $\breve{\mathcal{I}}_0 = 0_{p\times p}$, $\breve{\beta}_{n_0,0} = 0_{p\times 1}$, and $\breve{\mathcal{I}}_k = \sum_{i=1}^{k} \breve{\mathcal{I}}_{n_i,i}$. Estimator (3.8) is the weighted combination of the previous intermediary estimators $\breve{\beta}_{n_i,i}, i = 1, \ldots, k-1$ and the current subset estimator $\widehat{\beta}_{n_k,k}$. It results as the solution to the estimating equation $\sum_{i=1}^{k-1} \check{\mathcal{I}}_{n_i,i}(\beta - \check{\beta}_{n_i,i}) + \widehat{\mathcal{I}}_{n_k,k} \times (\beta - \widehat{\beta}_{n_k,k}) = 0$, with $\widehat{\mathcal{I}}_{n_k,k}(\beta - \widehat{\beta}_{n_k,k})$ being the bias correction term since $-\sum_{i=1}^{k-1} U_{n_i,i}$ has been omitted.

With $\check{\beta}_{n_k,k}$ given in (3.8), our CUEE estimator $\widetilde{\beta}_k$ is

$$\widetilde{\beta}_k = \left\{\check{\mathcal{I}}_{k-1} + \check{\mathcal{I}}_{n_k,k}\right\}^{-1} (s_{k-1} + \check{\mathcal{I}}_{n_k,k}\check{\beta}_{n_k,k} + \xi_{k-1} + U_{n_k,k}(\check{\beta}_{n_k,k}))$$

with $s_k = \sum_{i=1}^{k} \check{\mathcal{I}}_{n_i,i}\check{\beta}_{n_i,i} = \check{\mathcal{I}}_{n_k,k}\check{\beta}_{n_k,k} + s_{k-1}$ and $\xi_k = \sum_{i=1}^{k} \check{U}_{n_i,i} = \check{U}_{n_k,k} + \xi_{k-1}$, where $s_0 = \xi_0 = 0_{p\times 1}$, and $k = 1, 2, \ldots$. For the variance of $\widetilde{\beta}_k$, as $0_{p\times 1} = -\widehat{U}_{n_k,k} \approx -\check{U}_{n_k,k} + \widehat{\mathcal{I}}_{n_k,k}(\widehat{\beta}_{n_k,k} - \check{\beta}_{n_k,k})$, we have $\check{\mathcal{I}}_{n_k,k}\check{\beta}_{n_k,k} + \check{U}_{n_k,k} \approx \check{\mathcal{I}}_{n_k,k}\widehat{\beta}_{n_k,k}$. The estimated variance of $\widetilde{\beta}_k$ is online updated by

$$\widetilde{\text{Var}}(\widetilde{\beta}_k) = \left(\check{\mathcal{I}}_{k-1} + \check{\mathcal{I}}_{n_k,k}\right)^{-1} \left(\check{\mathcal{I}}_{k-1}\widetilde{\text{Var}}(\widetilde{\beta}_{k-1})\check{\mathcal{I}}_{k-1}^{\top} + \check{\mathcal{I}}_{n_k,k}\widehat{\mathcal{I}}_{n_k,k}^{-1}\check{\mathcal{I}}_{n_k,k}^{\top}\right) \left[\left(\check{\mathcal{I}}_{k-1} + \check{\mathcal{I}}_{n_k,k}\right)^{-1}\right]^{\top}.$$

Upon further simplification, it reduces to

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

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).

3.2.4 Asymptotic Results

We now provide the asymptotic distribution of the test statistic $T_k(G)$ given in Equation (3.1). For ease of presentation, we assume that all subsets of data are of equal size n, i.e., $n_k = n$. 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 (1982).

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

C3 Assume that $\{X(t), t \in [0, \tau]\}$ 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 $\widehat{\beta}_{n,k}$ satisfies $P(n^{\alpha}\|\widehat{\beta}_{n,k} - \beta_0\| > \eta) \leq C_{\eta} n^{2\alpha-1}$, where $C_{\eta} > 0$ is a constant only depending on η.

C5 For each subdata, $\|\sum_{\ell=1}^{d_k} G_{\ell k}\widehat{V}_{\ell k}\| < C_{gv} n \|\widehat{\beta}_{n,k} - \beta_0\|$, or $\|\sum_{\ell=1}^{d_k} G_{\ell k}\check{V}_{\ell k}\| < C_{gv} n \|\check{\beta}_{n,k} - \beta_0\|$, where C_{gv} is a constant that does not depend on k.

The conditions assumed in Section 2.4 of Andersen (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 varying covariate. If it is time invariant, 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 $\| \sum_{\ell=1}^{d_k} G_{\ell k} \widehat{V}_{\ell k} \| = O_P(\sqrt{n})$. This condition is typically satisfied in practice. As mentioned in Therneau and Grambsch (2000), $\widehat{V}_{\ell k}$ are often replaced by $\mathcal{I}_{n_k,k}/d_k$ in practice and $G_{\ell k}$ are always centered. Thus, $\sum_{\ell=1}^{d_k} G_{\ell k} \widehat{V}_{\ell k} = 0$ for this scenario.

Theorem 3.2.1. *Under conditions C1-C5, as $n \to \infty$, if $K = O(n^\gamma)$ with $0 < \gamma < \min\{1 - 2\alpha, 4\alpha - 1\}$, then for any $k \le K$, the test statistic satisfies that*

$$T_k(G) \to \chi_p^2,$$

in distribution when all blocks of data follow the proportional hazards model with the same covariate parameters.

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

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

$$\Gamma_K = H_K^{-1/2} Q_K, \quad \text{where} \quad H_K = \sum_{k=1}^{K} \sum_{\ell=1}^{d_k} G_{\ell k} \widehat{V}_{\ell k} G_{\ell k}. \tag{3.9}$$

To prove the asymptotic chi-square distribution, we only need to show that Γ_K converges in distribution to a p-dimensional multivariate standard normal distribution.

We first show that $(nK)^{-1} H_K$ 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

$$\frac{1}{n}\sum_{\ell=1}^{d_k} G_{\ell k}\widehat{V}_{\ell k}G_{\ell k} \to \int_0^\tau G(t)v(\beta_0,t)G(t)s^{(0)}(\beta_0,t)\lambda_0(t)\mathrm{d}t \equiv \Sigma, \tag{3.10}$$

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

Since $\{X(t), t \in [0,\tau]\}$ is a bounded Donsker class, $\{Y(t)\exp\{\beta' X(t)\}, t \in [0,\tau], \beta \in \mathbb{B}\}$ is also Donsker. A Donsker class is also a Glivenko-Cantelli class, so we have

$$\sup_{t\in[0,\tau],\beta\in\mathbb{B}}\left|\frac{1}{n}\sum_{0}^{n} Y_{\ell k}(t)\exp\{X_{\ell k}(t)^\top\beta'\} - s^{(0)}(\beta,t)\right| \to 0, \tag{3.11}$$

almost surely, where $\mathbb{B}$ is the compact parameter space. This means that $S^{(0)}(\beta,t)$ is uniformly bounded away from 0. As a result, $\frac{1}{n}\sum_{\ell=1}^{d_k} G_{\ell k}\widehat{V}_{\ell k}G_{\ell k}$ is bounded since the covariate $X(t)$ is bounded. Thus, from Theorem 1.3.6 of Serfling (1980), Equation (3.10) implies that

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

With this, from Fubini's theorem, we have

$$E\left\{\frac{H_K}{nK}\right\} = \frac{1}{K}\sum_{k=1}^{K} E\left\{\frac{1}{n}\sum_{\ell=1}^{d_k} G_{\ell k}\widehat{V}_{\ell k}G_{\ell k}\right\} \to \Sigma.$$

Thus,

$$\frac{H_K}{nK} \to \Sigma, \tag{3.12}$$

in probability.

Now we examine $Q_K = \sum_{k=1}^{K} \sum_{\ell=1}^{d_k} G_{\ell k} \widehat{r}_{\ell k}$. For each component of $\widehat{r}_{\ell k}$, $\widehat{r}_{\ell k}^{(i)}$ $(i = 1, ..., p)$, the Taylor series expansion yields

$$\widehat{r}_{\ell k}^{(i)} = r_{\ell k}^{(i)} - V_{(i)}(\widehat{\beta}_{n,k}^{(i*)}, t_\ell)(\widehat{\beta}_{n,k} - \beta_0),$$

where $V_{(i)}(\widehat{\beta}_{n,k}^{(i*)}, t_\ell)$ is the ith row of $V(\widehat{\beta}_{n,k}^{(i*)}, t_\ell)$, and $\widehat{\beta}_{n,k}^{(i*)}$ is on the line segment between $\widehat{\beta}_{n,k}$ and β_0. If $V(\widehat{\beta}_{n,k}^{*}, t_\ell)$ is the matrix whose rows are $V_{(i)}(\widehat{\beta}_{n,k}^{(i*)}, t_\ell)$, $i = 1, ..., p$, then we have

$$\widehat{r}_{\ell k} = r_{\ell k} - V(\widehat{\beta}_{n,k}^{*}, t_\ell)(\widehat{\beta}_{n,k} - \beta_0).$$

Thus

$$\begin{aligned} Q_K &= \sum_{k=1}^{K} \sum_{\ell=1}^{d_k} G_{\ell k} \widehat{r}_{\ell k} \\ &= \sum_{k=1}^{K} \sum_{\ell=1}^{d_k} G_{\ell k} r_{\ell k} - \sum_{k=1}^{K} \sum_{\ell=1}^{d_k} G_{\ell k} V(\widehat{\beta}_{n,k}^{*}, t_\ell)(\widehat{\beta}_{n,k} - \beta_0) \equiv \Delta_1 - \Delta_2. \end{aligned} \tag{3.13}$$

Note that Δ_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

$$\frac{\Delta_1}{\sqrt{nK}} \to N(0, \Sigma), \tag{3.14}$$

in distribution. Now we show that

$$\frac{\Delta_2}{\sqrt{nK}} = o_P(1). \tag{3.15}$$

Note that for each k, $\| \sum_{\ell=1}^{d_k} G_{\ell k} \widehat{V}_{\ell k} \| < C_{gv} n \|\widehat{\beta}_{n,k} - \beta_0\|$. Thus,

$$\begin{aligned} \|\Delta_2\| \leq & \sum_{k=1}^{K} \left\| \sum_{\ell=1}^{d_k} G_{\ell k} \{V(\widehat{\beta}^*_{n,k}, t_\ell) - V(\widehat{\beta}_{n,k}, t_\ell)\}(\widehat{\beta}_{n,k} - \beta_0) \right\| \\ & + \sum_{k=1}^{K} \left\| \sum_{\ell=1}^{d_k} G_{\ell k} V(\widehat{\beta}_{n,k}, t_\ell)(\widehat{\beta}_{n,k} - \beta_0) \right\| \\ & \leq C_g \sum_{k=1}^{K} \sum_{\ell=1}^{d_k} \|V(\widehat{\beta}^*_{n,k}, t_\ell) - V(\widehat{\beta}_{n,k}, t_\ell)\| \|\widehat{\beta}_{n,k} - \beta_0\| + C_{gv} n \sum_{k=1}^{K} \|\widehat{\beta}_{n,k} - \beta_0\|^2, \end{aligned} \tag{3.16}$$

where C_g is a constant that bounds $G(t)$ from above.

For the $i_1 i_2$th element of $V(\widehat{\beta}^*_{n,k}, t_\ell) - V(\widehat{\beta}_{n,k}, t_\ell)$,

$$V_{(i_1 i_2)}(\widehat{\beta}^*_{n,k}, t_\ell) - V_{(i_1 i_2)}(\widehat{\beta}_{n,k}, t_\ell) = \frac{\partial V_{(i_1 i_2)}(\widehat{\beta}^{**}_{n,k}, t_\ell)}{\partial \beta} (\widehat{\beta}^*_{n,k} - \widehat{\beta}_{n,k}),$$

where $\widehat{\beta}^{**}_{n,k}$ is on the line segment between $\widehat{\beta}^*_{n,k}$ and $\widehat{\beta}_{n,k}$. From (3.11) and the fact that

$X(t)$ is bounded, we know that $\partial V_{(i_1 i_2)}(\widehat{\beta}^{**}_{n,k}, t_\ell)/\partial\beta$ is uniformly bounded. Let M be a constant that bounds its elements. Since $\widehat{\beta}^{**}_{n,k}$ and $\widehat{\beta}^{*}_{n,k}$ are between $\widehat{\beta}_{n,k}$ and β_0, we have

$$|V_{(i_1 i_2)}(\widehat{\beta}^{*}_{n,k}, t_\ell) - V_{(i_1 i_2)}(\widehat{\beta}_{n,k}, t_\ell)| \leq M\|\widehat{\beta}_{n,k} - \beta_0\|. \tag{3.17}$$

Combining (3.16) and (3.17), we have

$$\|\Delta_2\| \leq Cn \sum_{k=1}^{K} \|\widehat{\beta}_{n,k} - \beta_0\|^2, \tag{3.18}$$

where $C = C_g M + C_{gv}$. Since $K = O(n^\gamma)$, there exist a constant, say C_1^2, such that $K < C_1^2 n^\gamma$. From (3.18), for any $\epsilon > 0$,

$$\begin{aligned}
P\big(\|\Delta_2\| > \sqrt{nK}\epsilon\big) &\leq P\bigg(\frac{1}{K}\sum_{k=1}^{K}\|\widehat{\beta}_{n,k} - \beta_0\|^2 > \frac{\epsilon}{C\sqrt{nK}}\bigg) \\
&\leq \sum_{k=1}^{K} P\bigg(\|\widehat{\beta}_{n,k} - \beta_0\|^2 > \frac{\epsilon}{C\sqrt{nK}}\bigg) \\
&\leq \sum_{k=1}^{K} P\bigg(\sqrt{nn^\gamma}\|\widehat{\beta}_{n,k} - \beta_0\|^2 > \frac{\epsilon}{CC_1}\bigg) \\
&= \sum_{k=1}^{K} P\bigg(n^{(1+\gamma)/4}\|\widehat{\beta}_{n,k} - \beta_0\| > \sqrt{\frac{\epsilon}{CC_1}}\bigg) \\
&\leq \sum_{k=1}^{K} P\bigg(n^{\alpha}\|\widehat{\beta}_{n,k} - \beta_0\| > \sqrt{\frac{\epsilon}{CC_1}}\bigg) \\
&\leq \sum_{k=1}^{K} C_\eta n^{2\alpha-1} = C_\eta K n^{2\alpha-1} = O(n^{\gamma+2\alpha-1}) = o(1).
\end{aligned}$$

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 (3.15). The proof finishes by combining (3.9), (3.12), (3.13), (3.14), (3.15), and Slutsky's theorem.

Now we consider the case when $\mathcal{I}_{n,k}$ and $\widehat{r}_{\ell k}$ are evaluated at $\breve{\beta}_{n,k}$. 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 $\widehat{\beta}_{n,k}$ is also valid for $\breve{\beta}_{n,k}$. With this result, the proof is similar to the case when $\mathcal{I}_{n,k}$ and $\widehat{r}_{\ell k}$ are evaluated at $\widehat{\beta}_{n,k}$. □

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 n_k's are different, the asymptotic result is still valid under mild some condition, for example, $\max_k n_k / \min_k n_k = O(1)$. Note that the window version statistic $T_k^w(G)$ 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 $T(G)$ on the most recent block, which has been shown to be χ_p^2 by Grambsch and Therneau (1994).

3.3 Simulation Study

Simulation studies were carried out to evaluate the empirical size and power of both the online updating cumulative and window versions of the test statistic. When data were generated under the proportional hazards assumption, we also compared the empirical distribution of the online updating cumulative statistic $T_k(G)$ with that of the standard statistic computed using all data up to selective accumulation points k, denoted by $T_{1:k}(G)$. While we look at the end of each stream to decide whether the entire stream of data satisfies the proportional hazards 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 in memory usage for the proposed statistics.

3.3.1 Size

Event times were generated from Model (2.1) 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 $i = 1, \dots, n_k$, making a $n_k \times 3$ covariate matrix. We set a vector of parameters $\beta_0 = (0.67, -0.26, 0.36)^\top$, 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. 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 $n_k = 2,000$.

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 the cumulative

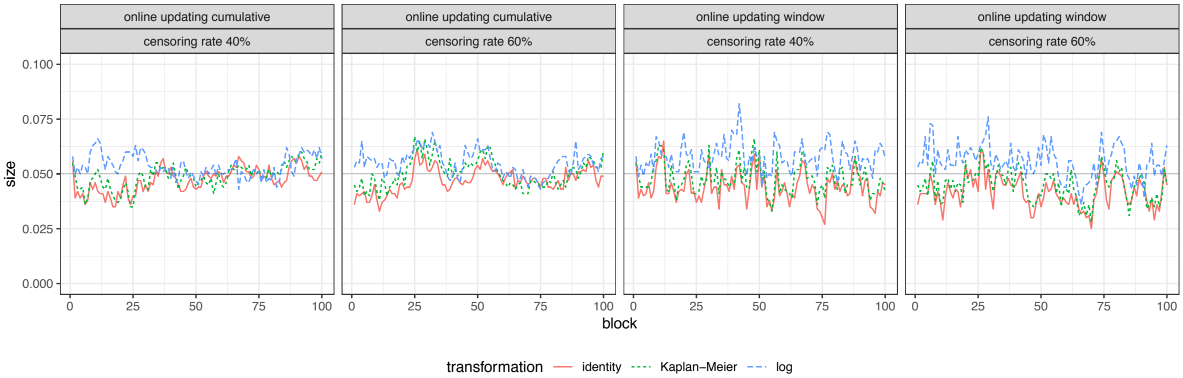

Figure 9: Empirical size (proportion of statistic values greater than $\chi^2_{3,0.95}$) calculated at each update using the identity, KM, and log transformations under the null hypothesis.

version and window version (width $w = 5$) statistics upon arrival of each block of simulated data. For the cumulative version statistic, the matrices and Schoenfeld residuals were evaluated at $\widetilde{\beta}_k$, the CUEE estimator. Figure 9 summarizes empirical rejection rates 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 rejection rates 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 than the other two transformations, and its usage should therefore be treated with caution.

To compare the empirical distribution of the online updating cumulative statistic $T_k(G)$ and the standard statistic $T_{1:k}(G)$, we generated $1,000$ independent streams of data, each again with $K = 100$ blocks and $n_k = 2,000$ under the same settings as before. Test statistics $T_k(G)$ were computed for all blocks $k = 1, \ldots, 100$ according to Equation (3.1). At blocks $k \in \{25, 50, 75, 100\}$, we also calculate the standard statistic $T_{1:k}(G)$ based on cumulative data up to those blocks; that is, we combine the data in

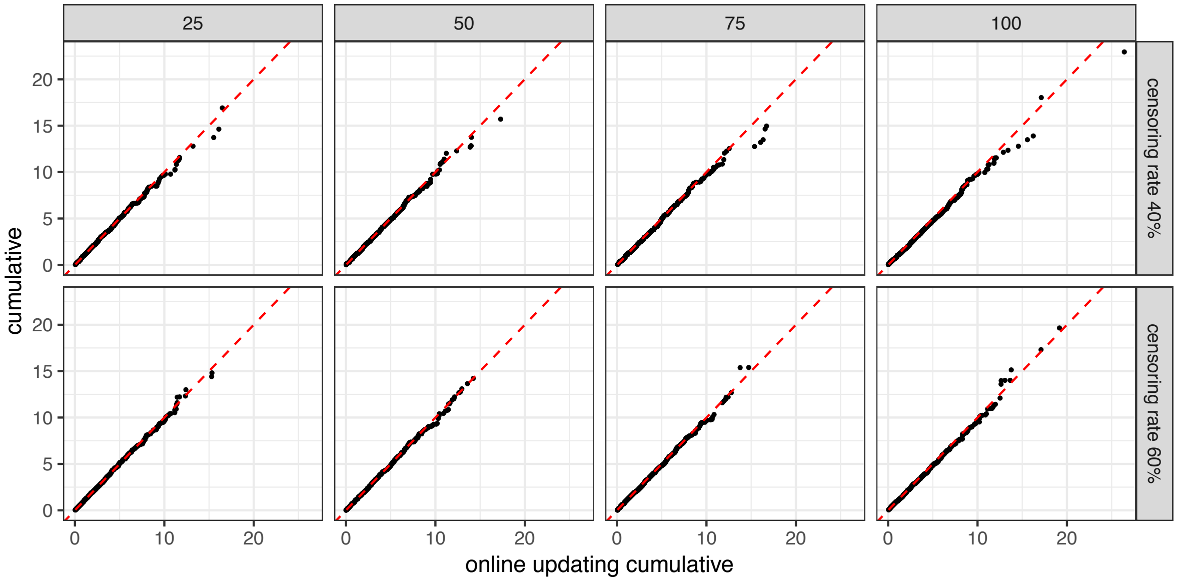

Figure 10: Empirical quantile-quantile plots of the online updating cumulative statistics $T_k(G)$ (x-axis) and $T_{1:k}(G)$ obtained using cumulative data (y-axis) with censoring rate 40% and 60%, taken at block $k \in \{25, 50, 75, 100\}$, both calculated using the KM transformation on event times.

block k with the previous $k-1$ blocks into a single large dataset and obtain $T(G)$ in Equation (2.10) based on this single large dataset of k blocks. Figure 10 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 $T_k(G)$ follow the same asymptotic χ^2_p distribution under the null hypothesis as $T_{1:k}(G)$.

In addition to scenarios where $p = 3$, simulation studies are performed to assess the size of the proposed test statistics for moderate dimensions for $p \in \{10, 20\}$. For each setting, there are $p/2$ continuous covariates, generated i.i.d. from $\mathcal{N}(0, 1)$, and the remaining $p/2$ covariates are binary, generated i.i.d. from Bernoulli(0.5). The beta

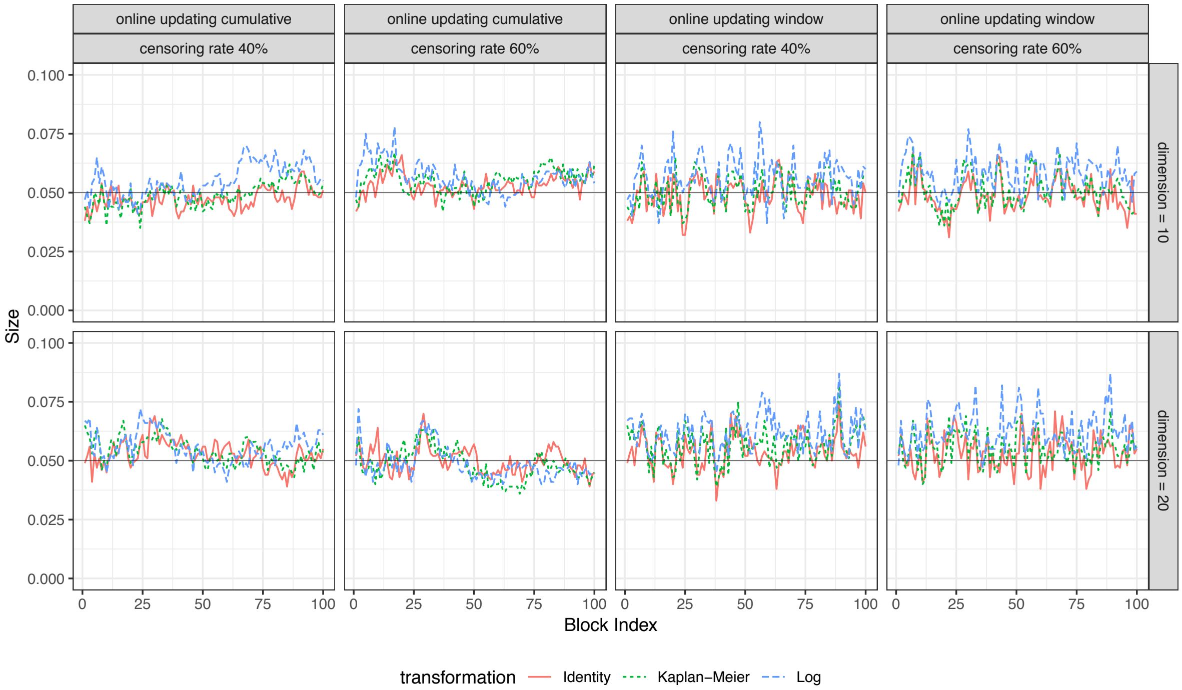

Figure 11: Size for the proposed test statistics when $p = 10$ and 20.

vectors are chosen as $\beta_{10} = (0.7, -0.5, 0.8, 0.3, 0.1, -0.4, -0.9, -0.2, -0.3, 0.4)^\top$, and $\beta_{20} = (\beta_{10}^\top, \beta_{10}^\top)$. The baseline hazards are set to, respectively, 0.032 and 0.015, with the weights at ⟨60⟩ being (0.9, 0.1) to produce the desired censoring rates of approximately 40% and 60%. It can be seen from Figure 11 that both versions of statistic hold their sizes under the null hypothesis, under both dimensions, although the log transformation is not recommended.

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 (see Section 3.4), we checked the size of the proposed test in a simulation study with a Cox model having a similar structure. The function survSplit() from the **survival**

Table 4: Size of $T_k(G)$ for models with piecewise constant coefficients based on 1,000 replicates.

Censoring Rate	Transformation	Size
40%	KM	0.067
	Identity	0.043
	Logarithm	0.156
60%	KM	0.039
	Identity	0.033
	Logarithm	0.094

package 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. (2017). As an illustration, we used the **reda** package (Wang et al., 2017) to simulate survival data with again the three covariates, but the coefficients are now piecewise constant. On the interval $[0, 12]$, $\beta = (0.7, -0.26, 0.36)$, and on the interval $(12, 60]$, $\beta = (0.6, -0.4, 0.46)$. The same censoring schemes as in earlier this section 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 $T_k(G)$ 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 Table 4. 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 KM or identity transformations are used. The logarithm transformation does not maintain its size well, which is similar to the instability we observed in Figure 9 and Figure 11, and is again not recommended.

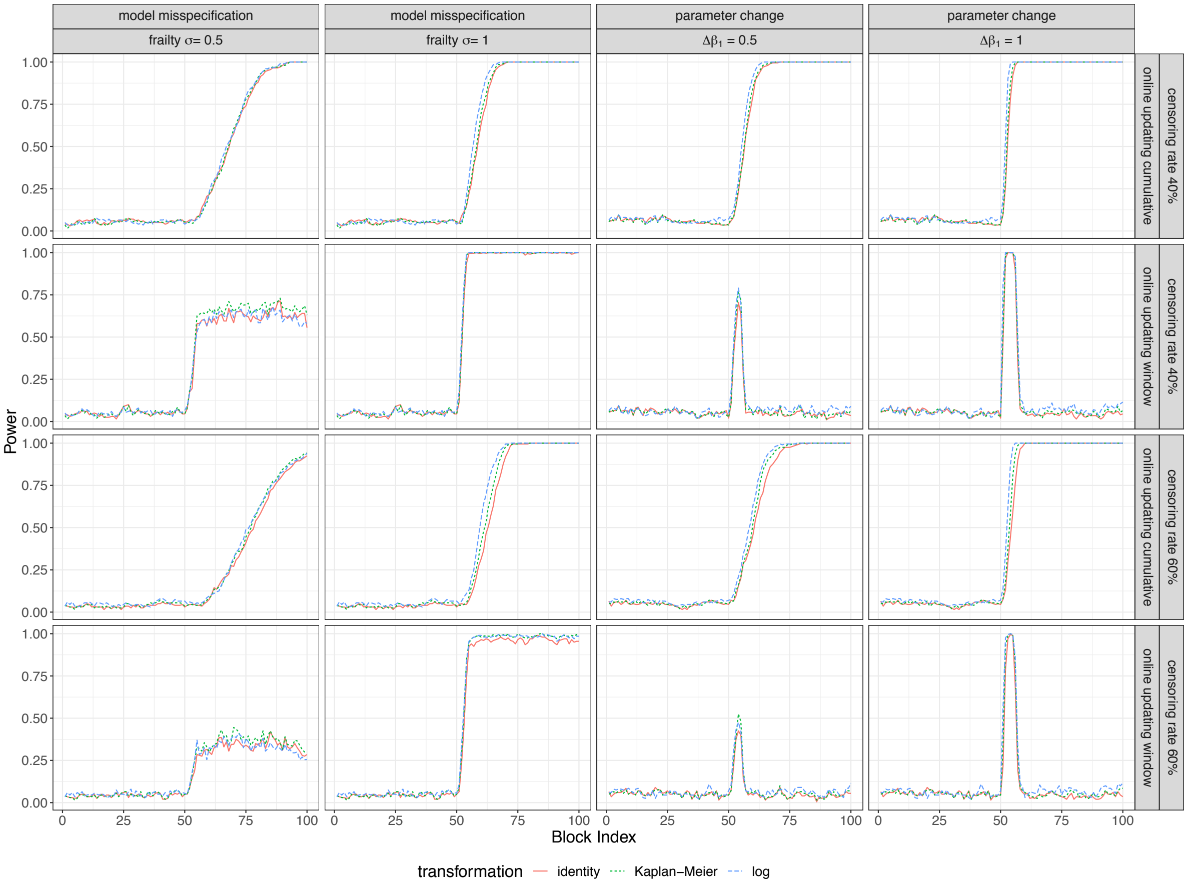

Figure 12: Empirical power (proportion of statistic values greater than $\chi^2_{3,0.95}$) for the online updating cumulative and window tests, calculated at each update using the identity, Kaplan–Meier, and log transformations under the alternative hypotheses of model misspecification (left) and parameter change (right) under censoring rate 40% (top) and 60% (bottom).

3.3.2 Power

Continuing with the simulation setting, two scenarios where the proportional hazards assumption is violated were considered to assess the power of the proposed tests.

The first scenario breaks the proportional hazards assumption by a multiplicative frailty in the hazard function. Starting from the 51st block in each stream, the hazard function, instead of being (2.1), becomes

$$\lambda(t) = \lambda_0(t) \exp\bigl(X^\top \beta + \epsilon\bigr),$$

where a normal frailty $\epsilon \sim N(0, \sigma^2)$ is introduced. Two levels of σ were considered, 0.5 and 1. Figure 12 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, the window version picks up the change more rapidly than the cumulative version because it discards information from older blocks for which the proportional hazards 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. The cumulative version responds to the change more slowly, but 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 transformation and KM transformation seem to have higher power than that based on the identity transformation.

The second scenario breaks the proportional hazards assumption by a change in one of the regression coefficients. 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 12. 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, the window version only has power to detect the change near the 51st block, when the blocks in the window contain data from both the original model and the changed model. The cumulative version picks up the change after the 51st block and the power increases quickly to 1 as more data blocks from the changed model accumulate.

To further compare the powers of $T(G)$ and $T_k(G)$, in both scenarios, we decreased the magnitude of change in the underlying model generating the data streams, and calculate the powers of $T(G)$ and $T_k(G)$ at the end of each stream. 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 $T(G)$ and $T_k(G)$. 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 Figure 13.

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 $T(G)$ and $T_k(G)$ identify the violation with quite high power. The performance of $T_k(G)$ is not better than, but still comparable to, the performance of $T(G)$. Note that both statistics have higher power for the same change at 40% censoring level than at 60% censoring level.

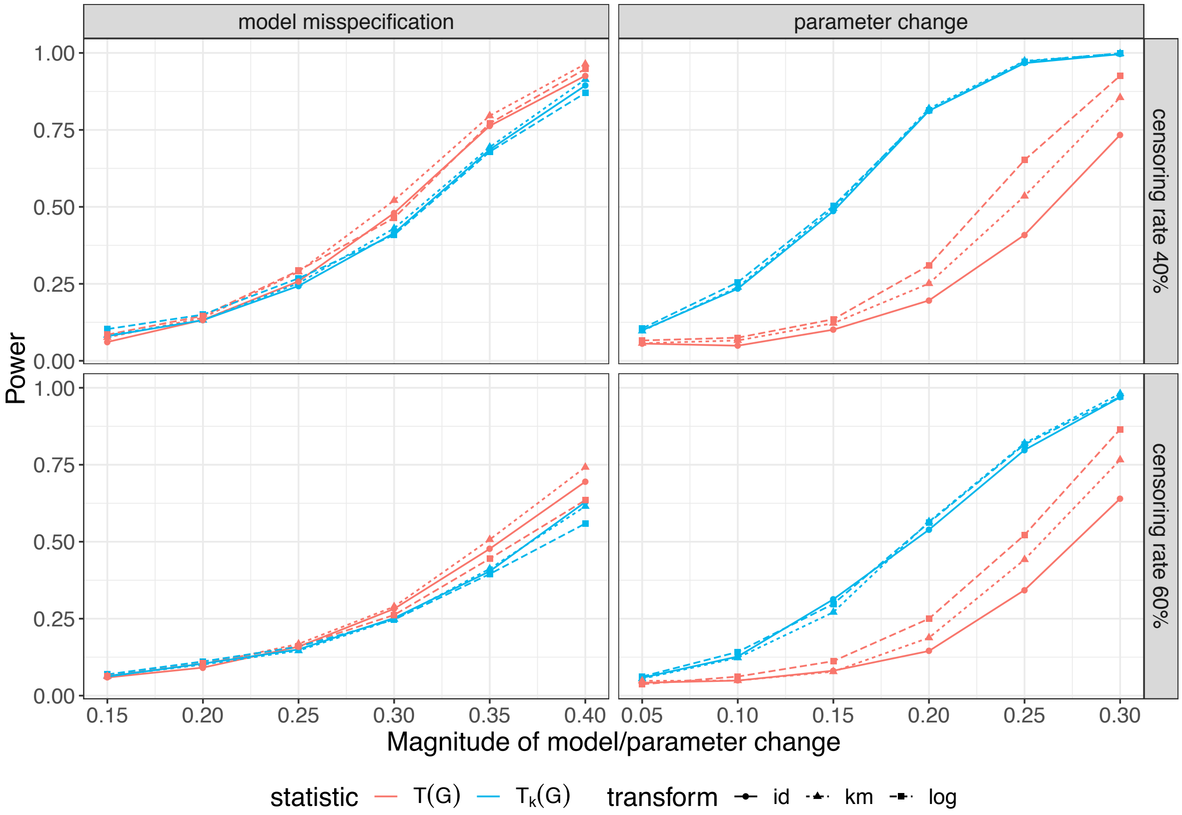

Figure 13: Power of $T(G)$ and $T_k(G)$ calculated at the end of the data stream when a violation occurs at the 51st block in each stream, plotted against the magnitude of violations. For the model misspecification scenario, the x axis denotes the frailty standard deviation, σ; for the parameter change scenario, the x axis denotes the change in β_1, i.e., $\Delta\beta_1$.

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

3.3.3 Comparison of Computing Time

The computation time for the standard test $T(G)$ and the online updating cumulative statistic $T_k(G)$, for both the CEE- and CUEE-based versions. For comparison purposes, we choose to simulate data that can be loaded into one computer's memory. Survival data streams using the setting of Section 3.3.1 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 $K = 100$ 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 β and the diagnostic statistic $T(G)$ are recorded, as well as the time it takes to obtain $T_k(G)$, $\widehat{\beta}_k$ and $\widetilde{\beta}_k$ 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 average computing time in Figure 14. 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 $T_k(G)$ is slightly slower than the CEE-based

$T_k(G)$, but the difference is minor. Note that $T(G)$ is only computed at the end of each stream. If we want to obtain a new $T(G)$ on cumulative data upon the arrival of each new block, like we can do with $T_k(G)$, the contrast of computing time would be even more significant.

To compute $T(G)$ on the entire data stream with N observations and d events, we first need to evaluate the log partial likelihood (2.3). 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 (2.10) 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 big dataset is partitioned into. Note, however, that K needs to satisfy the regularity condition in

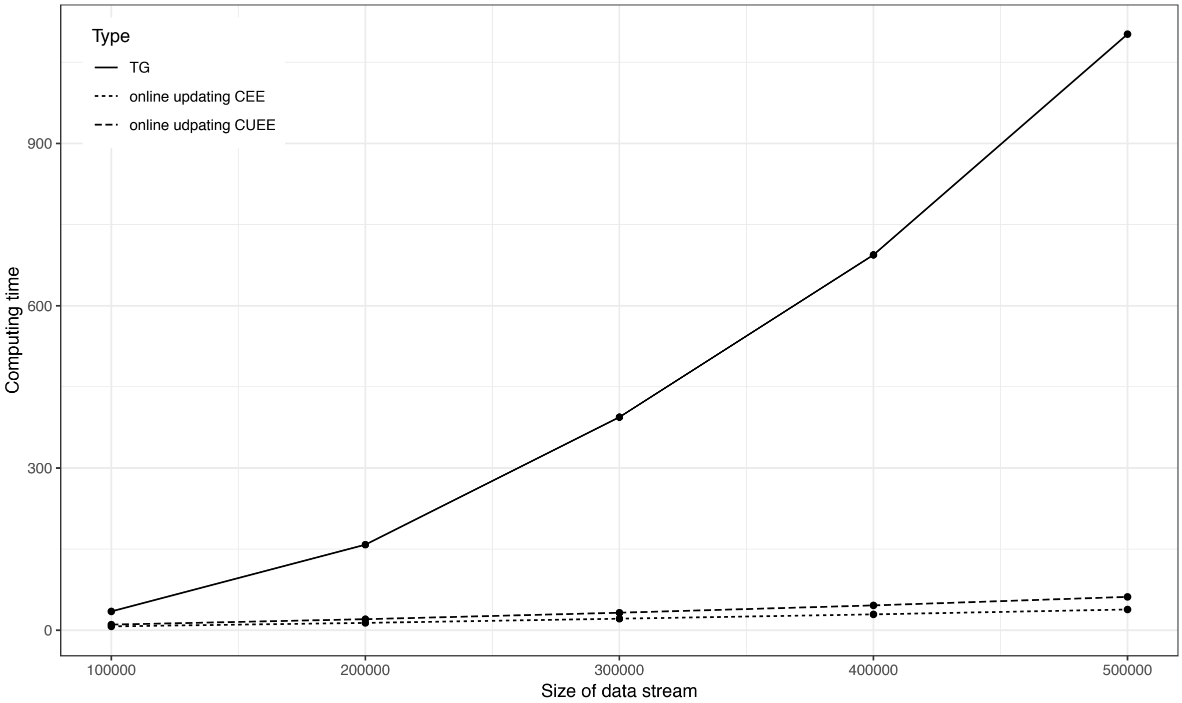

Figure 14: Plot of average computing times over 100 replicates for $T(G)$ and two versions of $T_k(G)$ when stream size $n \in \{100000, 200000, 300000, 400000, 500000\}$ and each stream is partitioned into 100 equally sized blocks.

Theorem 3.2.1.

3.3.4 Comparison of Memory Usage

In addition to computing time, we also study the savings in memory usage of our proposed online updating statistics. A big dataset was simulated using the parameter setting in Section 3.3.1 with $\beta = (0.67, -0.26, 0.36)$ and $\lambda_0(t) = 0.018$, which contains $N = 200$ millions of 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 Section 3.3.3. The total memory available on this laptop is 32 GB. The **profvis** package (Chang and Luraschi, 2018) was used to track the 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 is 16,785.2 MB. Next, the online updating CUEE-based $T_k(G)$ 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. 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 load the entire dataset into R's workspace. The read.csv() procedure did not finish after running for an hour and occupying 37.57 GB of the virtual memory and 15.61 GB of the real memory, and finally aborted because of insufficient memory.

3.4 Survival Analysis of SEER Lymphoma Patients

We consider analyzing the survival time of the lymphoma patients in the SEER program with the proposed methods. There were 131,960 patients diagnosed with lymphoma cancer between 1973 to 2007. We limited our scope to events due to lymphoma within the first 60 months after being diagnosed Among those 131,960 subjects, the total number of events was 47,009, and the censoring rate was 64.4%. The risk factors considered in our analysis were Age (centered and scaled), gender indicator (Female), and African-American indicator (Black). There were 60,432 females, and 9,199 African-American. While the dataset is large, the analysis of the data as a single dataset is still possible with reasonable computing resources. We wish to compare the performance of the standard statistic $T(G)$ from Equation (2.10) with our online updating statistics under a setting in which the proportional hazards statistic is judged to be satisfied based on the standard $T(G)$ test. For online updating, the patients in the data were ordered by time of diagnosis, so it is natural to partition the data 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. Figure 15 presents the stacked bar plot of censors and events, together with the line plot of censoring rate for each block.

As a starting point, an initial model that included the three risk factors was fitted, and the standard test statistic based on the full data as in Equation (2.10) was calculated to be 83.38, which indicated that the model does not satisfy the proportional hazards

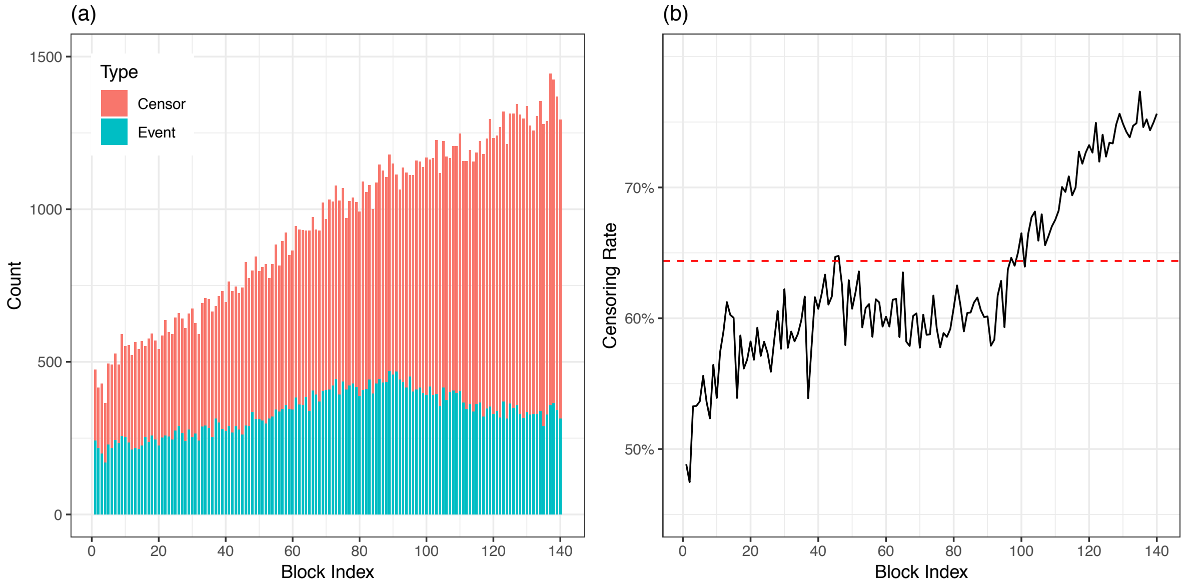

Figure 15: Sample size and censoring rate in blocks of SEER lymphoma data.

assumption. The online updating cumulative statistic was calculated to be 95.60. Due to the relatively high censoring rate, all diagnostics were applied after applying the Kaplan–Meier transformation on the survival times 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 Figure 16.

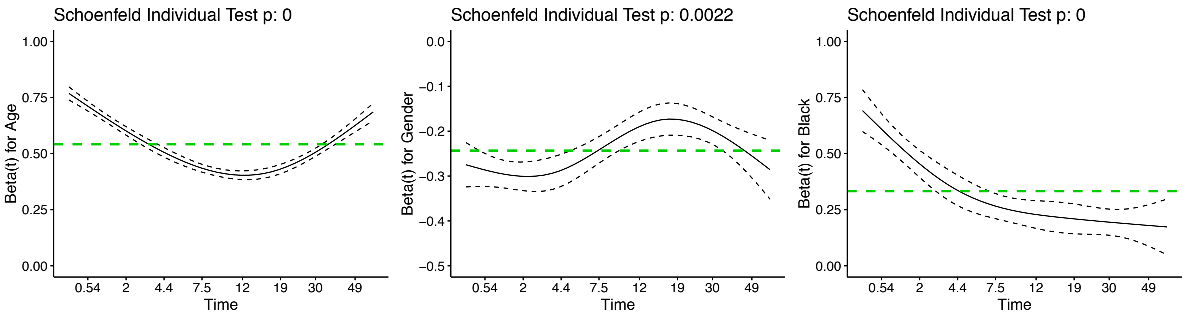

Figure 16: Time-varying pattern of the parameters for `Age`, `Gender` and `Black` in the initial model, with parameter estimates from the entire data overlaid in green.

Techniques in Therneau et al. (2017) were used to approximate the parameters using piecewise constant functions of time. Two cut-offs were chosen at 2 and 30 months based on the time-variation pattern obtained from the naive model. A factor variable tgroup is defined to indicate on which intervals the corresponding observation contributes to estimation of β. For example, a subject with survival time 25 and event 1 will now be represented separately on two intervals: one with time interval from 0 to 2, with event 0 and tgroup=1, and the other with time interval from 2 to 25, with event 1 and tgroup=2. The interaction of Age, Female and Black with the generated tgroup as strata, respectively, gives the model more flexibility to fit to the data. The new model resulted in $T(G) = T_{1:140}(G) = 5.75$ on 9 degrees of freedom with a p-value of 0.77, which indicates that the proportional hazards assumption for the revised model is appropriate based on the full data. Figure 17 presents time-dependency plot of parameters for the revised model. In contrast to Figure 16, 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.

To evaluate the performance of the online updating parameter estimates and test statistics under the revised model, at each block k, $k = 1, \dots, 140$, we calculated the parameter estimates, the online updating cumulative statistics $T_k(G)$, the online updating window version statistics $T_k^w(G)$, and also $T_{1:k}(G)$ based on the single large dataset consisting of all cumulative data up to block k. Two online updating cumulative statistics $T_k(G)$ were obtained, one using the CEE estimator $\widehat{\beta}_k$ and the other using the CUEE

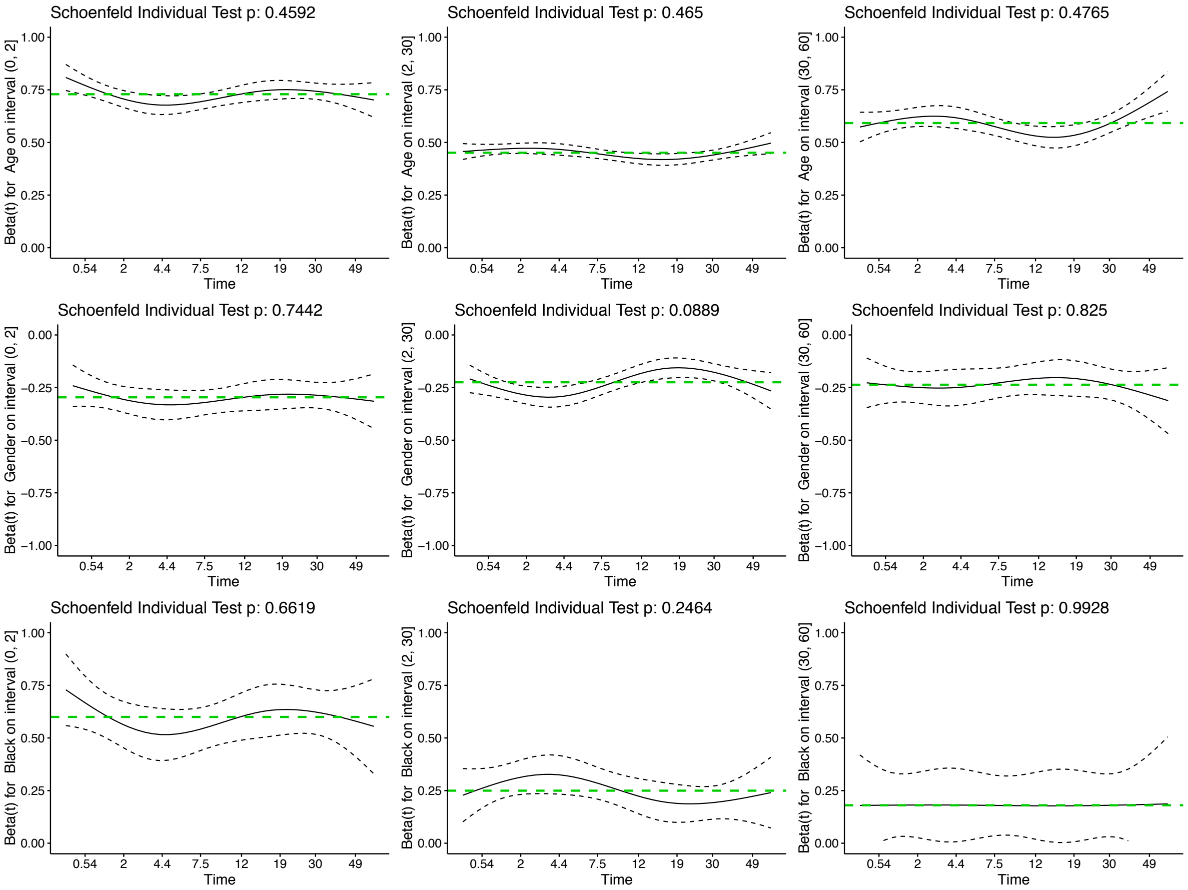

Figure 17: Time-varying pattern of the parameters for `age`, `Gender` and `Black` in the revised model on three disjoint intervals of survival time, with parameter estimates from the entire data overlaid in green.

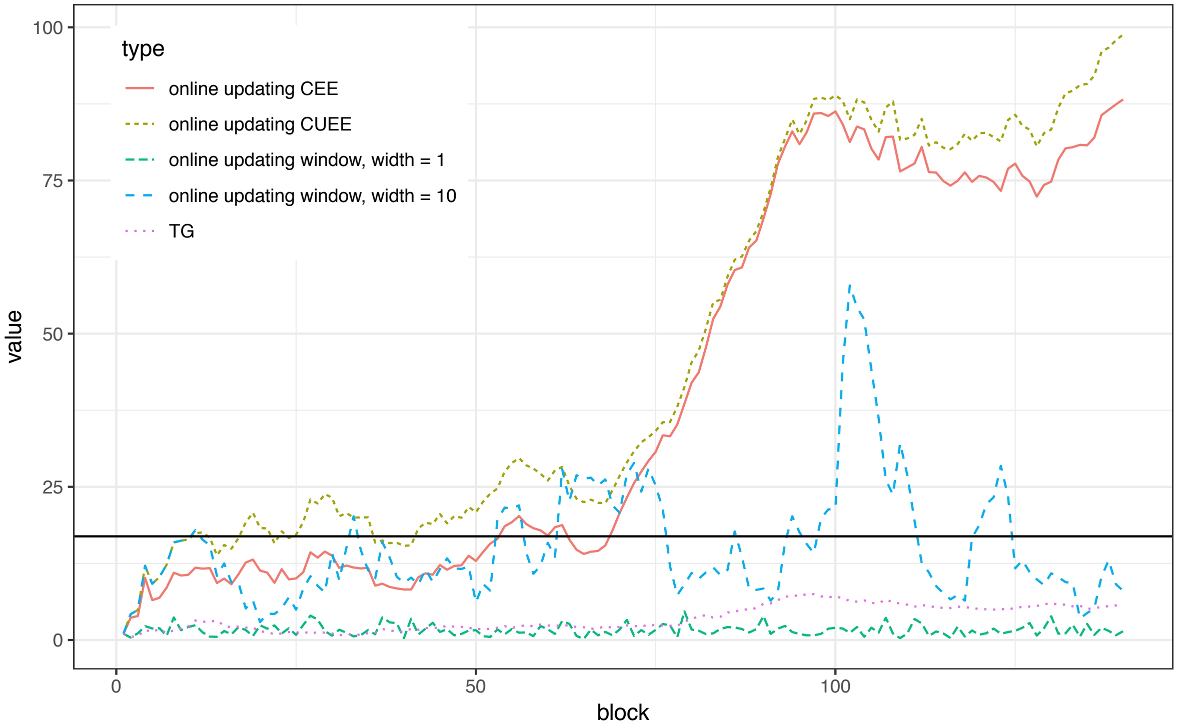

Figure 18: Test statistics for the proportional hazards hypothesis for lymphoma data, using temporally ordered dataset.

estimator $\widetilde{\beta}_k$. For the window version, the CEE estimator $\widehat{\beta}_k$ was used for computational convenience, and two widths $w = 1$ and $w = 10$ were considered. The trajectories of different versions of the test statistics were plotted in Figure 18. While the proportional hazards 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 $T_k(G)$ resulted in a rejection of the null hypothesis, and $T_k^w(G)$ when $w = 10$ also showed a few rejections along the stream.

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

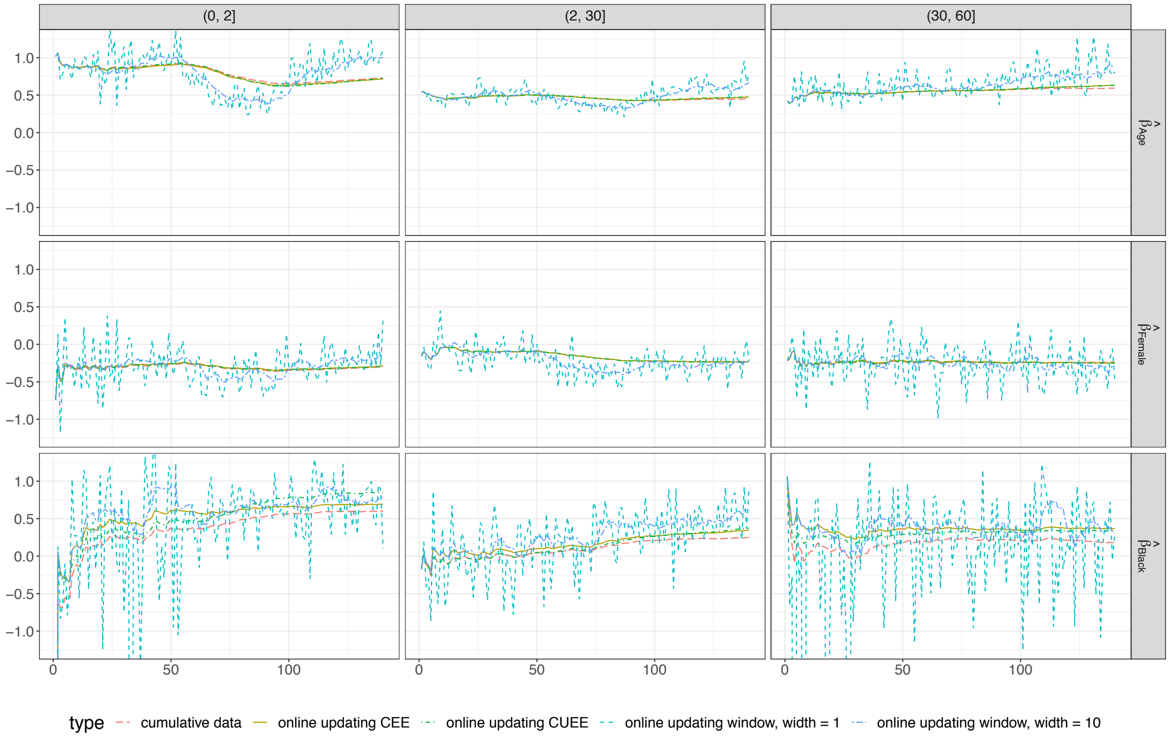

Figure 19: Parameter estimates given by different estimating schemes plotted with respect to block indices, obtained using the lymphoma data ordered by diagnosis time.

tgroup) were plotted with respect to block indices to investigate this apparent discrepancy; see Figure 19. Apparently, $\widehat{\beta}_{\mathsf{Age}}$ on $(0, 2]$ remained relatively stable for blocks 1 to 50, but started to first decrease and increase after. This change was captured by both the window ($T_k^w(G)$) and the cumulative version statistics ($T_k(G)$), but it was not captured by $T_{1:k}(G)$. This is explained by the fact that $T_{1:k}(G)$ is based on a single estimator of β, while in the online updating statistics, each block has its own estimate of β. The temporal changes that are observed in the CUEE estimate of β get canceled in the calculation based on the full cumulative data.

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 temporally-ordered, 3-month blocked data. For each permutation, we applied the same techniques and cutoffs to allow for piecewise constant parameters over time as before. The histogram of online updating cumulative statistics obtained for 1,000 such permutations is presented in Figure 20. 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 21 presents the same diagnostic plots as Figure 19 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 Figure 22. Each block again satisfies the proportional hazards assumption, and the performance of the online updating cumulative statistic based on CUEE is very close to $T(G)$ 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.

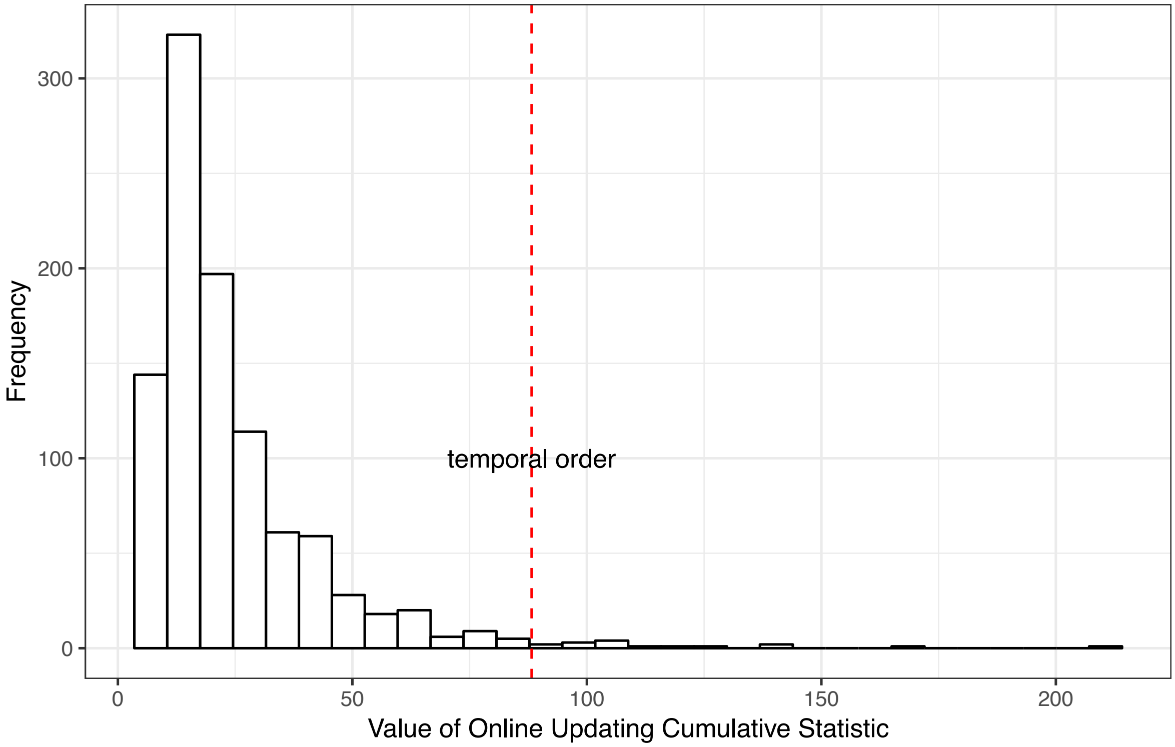

Figure 20: Histogram of online updating cumulative statistic obtained at the final block for 1,000 permutations of the original data, with observed test statistic value for the original data overlaid in red.

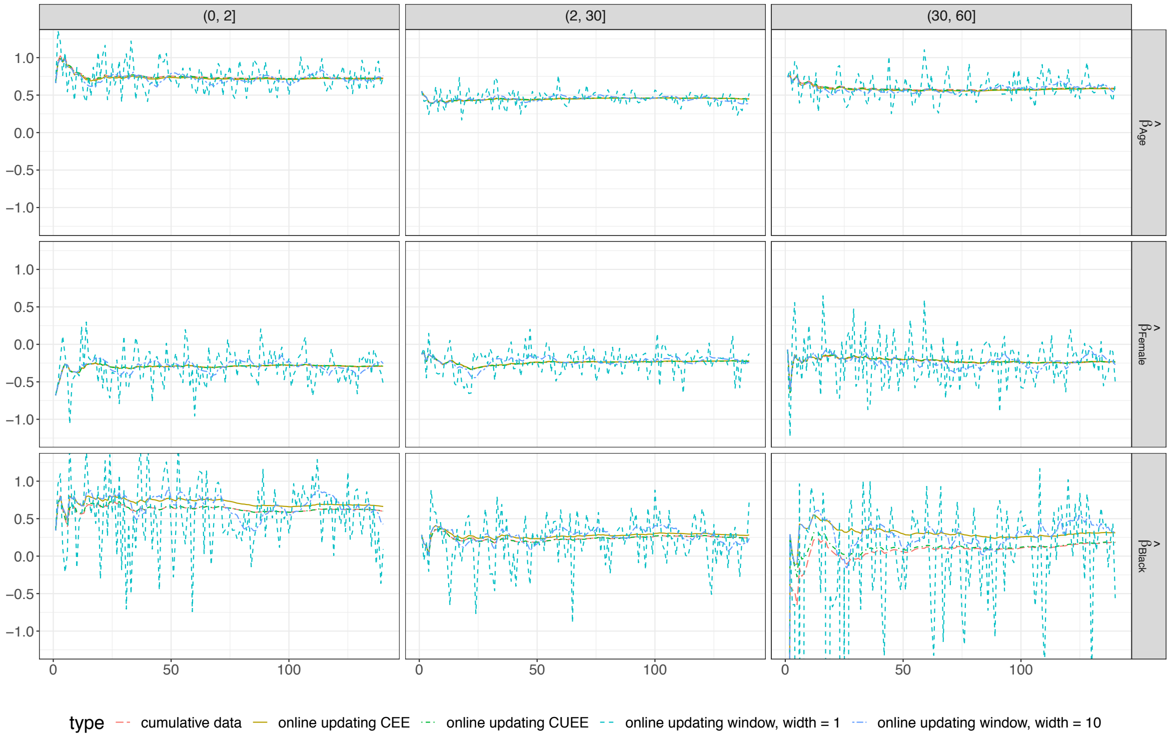

Figure 21: Parameter estimates given by different estimating schemes plotted with respect to block indices, obtained using the randomly ordered lymphoma data.

3.5 Discussion

We focus on the test for the proportional hazards assumption in this chapter. Specifically, instead of working on a given, full dataset, we developed the online updating test statistic for big streams of survival data. The statistic is 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). Two versions of test statistic was proposed: $T_k(G)$ that uses cumulative information from all historical data, and an auxiliary $T_k^w(G)$ using information only from recent data. Both statistic have an asymptotic χ^2 distribution when the blocks in the entire stream are

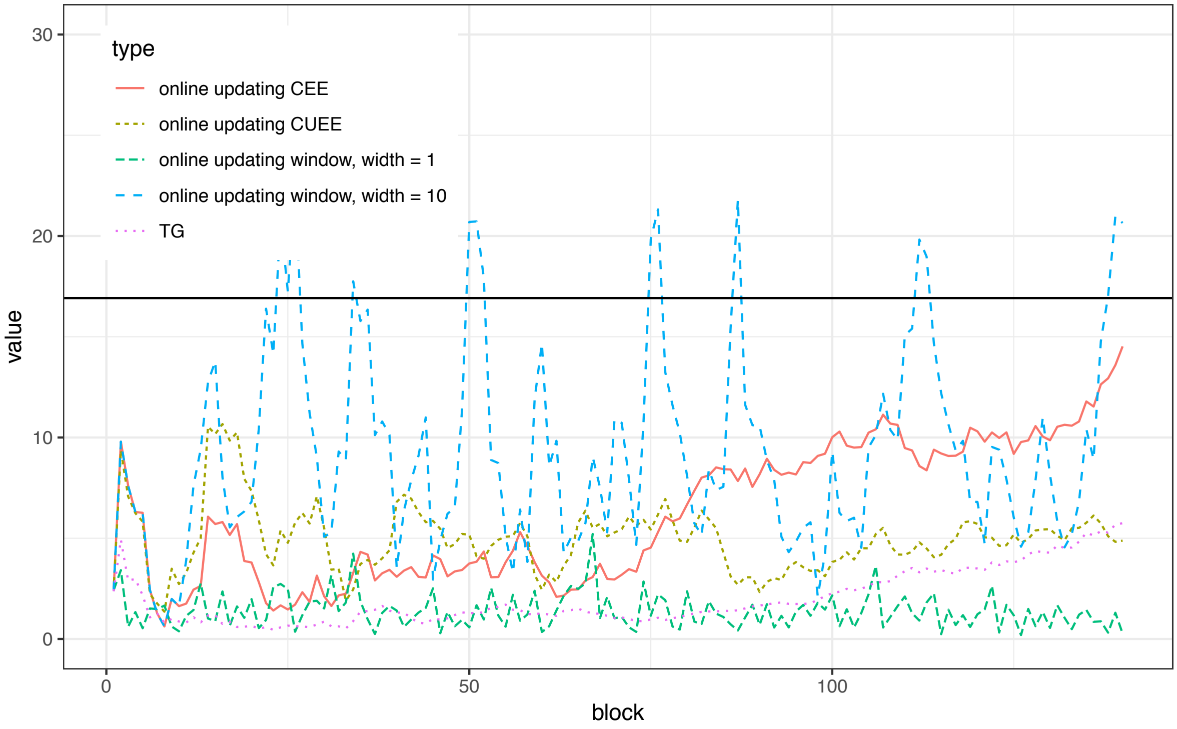

Figure 22: Test statistics for the proportional hazards hypothesis for lymphoma data, after the observations are randomly ordered.

generated from the same underlying Cox model. In the simulation studies, $T_k(G)$ has comparable or higher power to the standard test $T(G)$ of Grambsch and Therneau (1994) on the entire dataset, for scenarios of a model change or parameter change, respectively. In addition, when $T(G)$ fails to detect violation of the null hypothesis on the whole dataset, $T_k(G)$ may still identify the violation with high power. This was observed in the 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 that huge, it might be helpful 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 proportional hazards 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.

Compared to the traditional approach, $T_k(G)$ and $T_k^w(G)$ are computationally fast, and minimally storage intensive. Even when the dataset is too large to be loaded into the memory, the proposed approach can still be performed within reasonable time limit. Compared to parallel computing, 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 chapter are worth 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 $T_k^w(G)$, 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 12, $T_k^w(G)$ can behave differently under different violations of the proportional hazards 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 proportional hazards 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 et al., 2008; Fan et al., 2010; Mittal et al., 2014).

Also, we are only concerned with making a final decision regarding the proportional hazards 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 beings 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 proportional hazards test context are devoted for future research.

Chapter 4

Simultaneous Monitoring for Regression Coefficients and Baseline Profile in Cox Modeling

4.1 Introduction

The nonparametric baseline hazard function in the Cox model can be of special interest in applications where its change needs to be detected. Breslow (1972) proposed an approach, later summarized by Lin (2007) to be a nonparametric maximum likelihood estimation (NPMLE) approach, which estimates the regression parameters and the cumulative baseline hazard function at the same time. The resulting estimator of cumulative baseline hazard, when plotted against event times, form a monotone nondecreasing curve. In applications where the full survival curve of a given covariate set is needed for prediction, this nonparametric baseline hazard could also be of special interest. As a consequence, with a stream of survival data, changes in both the covariate effects and the baseline cumulative function need to be monitored.

Statistical profile monitoring techniques have been developed to detect changes in parameter vectors, or more complicated parameter profiles. Most parametric monitoring methods are applications of Hotelling's T^2 statistic (Hotelling, 1931). Kang and Albin (2000) monitored linear profiles by running a regression and keeping track of the estimated intercepts and slopes using a multivariate T^2 chart. Zhu and Lin (2009) used the same regression technique, but focused on using a t statistic for only the estimated slope after centering both the independent and dependent variables. The use of Hotelling's T^2 statistic was later extended to monitoring coefficients obtained by a parametric nonlinear regression in cases where a response curve is studied by Williams et al. (2007). Kazemzadeh et al. (2008) extended similar ideas to profiles obtained via a polynomial regression. In addition to the aforementioned parametric procedures, nonparametric monitoring methods have also been proposed. Zou et al. (2008) considered the use of nonparametric regression methods with some degree of smoothness in monitoring profiles. Woodall (2007) reviewed and summarized the application of such techniques in fields other than industrial manufacturing, including detecting changes in Q-Q plot reflecting the relationship between a collected sample and a baseline sample (Wang and Tsung, 2005) and detecting increased disease rate clusters (Zhou and Lawson, 2008). In an effort to allow the measurements within one profile to be correlated instead of strictly independent, Qiu et al. (2010) proposed the usage of nonparametric mixed effects models in profile monitoring context. Yu et al. (2012) formulated the profile monitoring problem in the scope of functional data analysis, and proposed an outlier detection mechanism based on functional principal component analysis. Wei et al. (2012) developed a purely nonparametric approach, which estimates a reference profile, and then relies on three nonparametric statistics to describe departures from the reference profile. This approach can be applied to essentially any curve-type observations, as the control limits can be established using quantiles of summary statistics based on existing data.

We considering statistical profile monitoring of Cox modeling in time-to-event data analysis where both the regression coefficient vector and the cumulative baseline hazard function need to be monitored. Following the conventions in statistical profile monitoring, we assume that in the beginning of a data stream, the blocks are "in control", i.e., observations in all blocks follow a Cox model with the same set of parameters and baseline hazard. The relatively homogeneous and stable blocks are then denoted as Phase I. Based on these blocks, after controlling for the heterogeneity in blockwise sample size and censoring rate, a T^2 statistic is used to describe a reasonable range of variation for the covariate effects, while three nonparametric descriptive statistics similar to those in Wei et al. (2012) are constructed for the same purpose for the cumulative hazard function. A combined decision rule is proposed to select the thresholds for the four statistics, such that the empirical type I error rate in Phase I is properly controlled. In Phase II, the same statistics are computed for each new arriving block. By comparing the values of the statistics with their respective control limits, one is able to tell whether a block is outlying in terms of one or more of the four measurements.

The rest of this chapter is organized as follows: in Section 4.2, we briefly outline the basics of the Cox model and the nonparametric estimate of the cumulative baseline hazard function, are obtained; in Section 4.3, we present, respectively, the monitoring methods for the coefficient estimates and for the cumulative baseline hazard function, and how to combine them to produce an integrated result. Simulation studies are presented in Section 4.4, followed by an application to the same lymphoma dataset from the SEER registry in Section 4.5.

4.2 Nonparametric Estimators for the Cumulative Baseline Hazard

Cox (1972, 1975) proposed the partial likelihood (2.2), which facilitates the estimation of covariate effects without having to consider the nonparametric baseline hazard function. Breslow (1972) used another formulation where a full model likelihood that incorporates both β and $\lambda_0(t)$ is used. The joint likelihood is written as:

$$L(\beta, \Lambda_0) = \prod_{i=1}^{n} \left\{ \exp\{X_i(T_i)^\top \beta\} \lambda_0(T_i) \right\}^{\delta_i} \exp\left\{ - \int_0^{T_i} \exp\{X_i(t)^\top \beta\} \lambda_0(t) \mathrm{d}t \right\}. \quad (4.1)$$

Maximizing the joint likelihood (4.1) with respect to β and λ_0 under the restriction that the baseline hazard is piecewise constant between uncensored event times yields $\widehat{\beta}_n$, which is exactly the maximum partial likelihood estimator, and the Breslow estimator for cumulative baseline hazard function:

$$\widehat{\Lambda}_0(t) = \sum_{s \le t} \left[\frac{\mathrm{d}N(s)}{\sum_{i=1}^n \exp\left(X_i^\top \widehat{\beta}_n\right) Y_i(s)} \right]. \tag{4.2}$$

A consistent approximation for the variance of $\widehat{\Lambda}_0(t)$ has been provided by Andersen and Gill (1982). At time t, we have

$$\sum_{s \le t} \left[\overline{X}(\widehat{\beta}_n, s)\mathrm{d}\widehat{\Lambda}_0(s) \right]^\top \mathcal{I}(\widehat{\beta}_n)^{-1} \sum_{s \le t} \left[\overline{X}(\widehat{\beta}_n, s)\mathrm{d}\widehat{\Lambda}_0(s) \right]^\top + \sum_{s \le t} \frac{\mathrm{d}\widehat{\Lambda}_0(s)}{\sum_{j=1}^n Y_j(s) \exp\left(X_j^\top \widehat{\beta}_n\right)}, \tag{4.3}$$

where $\overline{X}(\widehat{\beta}_n, s)$ is the $\overline{X}$ defined in (2.5) evaluated at $\widehat{\beta}_n$ and s. In practice, the survfit.coxph() function in the **survival** package enables one to obtain the estimated cumulative hazard, together with its variance estimate. It takes a coxph object, and if fed with a new data.frame object having the same structure as block1, computes the predictive survival curves for each unique cohort of covariates in the new dataset. If no such new dataset is fed, it computes the predicted survival curves at "average observation" whose covariates equal the means of covariates in the original data. Both survfit() and basehaz() in **survival** when applied to the same coxph object, returns a scaled version of the Breslow estimator:

$$\widehat{\Lambda}_{0_{\text{default}}}(t) = \exp\left(\overline{X}^\top \widehat{\beta}_n\right) \sum_{s \le t} \left[\frac{\mathrm{d}N(s)}{\sum_{i=1}^n \exp\left(X_i^\top \widehat{\beta}_n\right) Y_i(s)} \right],$$

where the scaling constant is the risk score of the "average observation". Different blocks, however, have different overall covariate levels, which may make the cumulative baseline hazard curves returned by `survfit()` not comparable. Therefore, to unmask the true Breslow estimator from the influence of covariate levels, we use a benchmark observation whose covariates are all 0. This is also a desirable practice, as has been discussed by the package author, if there are binary covariates such as an indicator for gender, a value of 1/2 is not reasonable.

4.3 Simultaneous Monitorng

In this section, we assume that we have already had K blocks of Phase I survival data, which all come from the same Cox model. Based on each block k with sample size n_k, similar to in Chapter 3, we obtain its maximum partial likelihood estimate $\widehat{\beta}_{n_k,k}$, its corresponding observed partial information matrix $\mathcal{I}_k(\widehat{\beta}_{n_k,k})$, the Breslow estimator $\widehat{\Lambda}_{n_k,k}(t)$ over a time grid $0 < t_1, \ldots, t_G \leq \tau$ for a pre-speficied τ, and the corresponding variances of the Breslow estimator at these time points as in (4.3).

4.3.1 Monitoring the Coefficient Estimates

As discussed in Williams et al. (2007), for nonlinear regression models, parameter estimates are usually obtained by numerically maximizing the likelihood function. Define a precision-weighted average vector of estimated parameters, $\overline{\widehat{\beta}}_K$ as

$$\overline{\widehat{\beta}}_K = \left(\sum_{k=1}^{K} \mathcal{I}_k(\widehat{\beta}_{n_k,k}) \right)^{-1} \sum_{k=1}^{K} \mathcal{I}_k(\widehat{\beta}_{n_k,k}) \widehat{\beta}_{n_k,k},$$

where the observed information matrices, i.e., precisions, are used as weights to ensure that each $\widehat{\beta}_{n_k,k}$ contributes to the weighted average proportionally to the amount of the information in block k. Intuitively, the more events a block contains (by having a larger sample size or a smaller censoring rate), the closer its parameter estimate is to the true underlying β_0. Substituting $\overline{\widehat{\beta}}_K$ into Equation (5) of Williams et al. (2007) produces the sample-size and censoring-rate adjusted version of the Hotelling's T^2 statistic:

$$T_k^2 = (\widehat{\beta}_{n_k,k} - \overline{\widehat{\beta}}_K)^\top S^{-1} (\widehat{\beta}_{n_k,k} - \overline{\widehat{\beta}}_K), \tag{4.4}$$

where S is some estimate of the covariance of $\widehat{\beta}_{n_k,k}$.

Several choices of S are possible. Williams et al. (2007) discussed the sample covariance matrix, one based on successive differences originally proposed by Holmes and Mergen (1993), an robust minimum volume ellipsoid (MVE) estimator proposed by Rousseeuw (1984), and a modified version of the MVE estimator. These estimates, however, all suffer from certain disadvantages, such as having doubtful power (Sullivan and Woodall, 1996), failing to be robust against multiple outliers (Vargas, 2003), or being computationally intensive and even hard to find (Jensen et al., 2007). In the Cox model scenario, fortunately, the observed information matrices are available. Denote $\mathcal{I}_k^{1/2}(\widehat{\beta}_{n_k,k})(\widehat{\beta}_{n_k,k} - \overline{\overline{\beta}}_K)$ as $\widetilde{\beta}_k$. Using similar argument as in Williams et al. (2007), when the number of events in block k is large enough, the distribution of $\widetilde{\beta}_k$ can be approximated by a multivariate normal distribution with mean $0_{p\times 1}$, and covariance matrix $I_{p\times p}$, which leads to the multivariate Hotelling's T^2 statistic:

$$\widetilde{T}_k^2 = \widetilde{\beta}_k^\top \widetilde{\beta}_k, \quad k = 1, \ldots, K. \tag{4.5}$$

For a Phase II block ℓ, denoting its vector of parameter estimates as $\widehat{\beta}^*_{n_\ell,\ell}$, we have its T^2 statistic:

$$\widetilde{T}_\ell^{2*} = (\widetilde{\beta}_\ell^*)^\top (\widetilde{\beta}_\ell^*), \quad \ell = 1, \ldots, \tag{4.6}$$

where $\widetilde{\beta}_\ell^* = \mathcal{I}_\ell^{*1/2}(\widehat{\beta}^*_{n_\ell,\ell})(\widehat{\beta}^*_{n_\ell,\ell} - \overline{\overline{\beta}}_K)$, where $\mathcal{I}_\ell^*(\widehat{\beta}^*_{n_\ell,\ell})$ is the observed partial information matrix for Phase II block ℓ, evaluated at $\widehat{\beta}^*_{n_\ell,\ell}$.

An appropriate control limit can be obtained by taking proper percentiles of the empirical distribution for the K Phase I statistics, $\widetilde{T}_1^2, \ldots, \widetilde{T}_K^2$, and based on the comparison of $\widetilde{T}_\ell^{2*}$ with the control limit, we are able to decide if the ℓth Phase II block differs significantly from being normative in terms of its covariate effects. Note that here we are making full use of the estimated covariance matrix of each $\widehat{\beta}_{n_k,k}$, so that inhomogeneity in blockwise sample sizes are accounted for, while in the existing approach of Williams et al. (2007), all profiles are assumed to be based on the same number of observations.

4.3.2 Monitoring the Cumulative Hazard Function

We consider a decomposition of a curve into a scalar center, a shape curve, and a variation curve similar to that in Wei et al. (2012). Extra challenges come from the fact that our cumulative hazard profiles are estimated instead of directly observed, and that differences in censoring rates and sample sizes across the sample blocks need to be taken into account. For each Phase I block k, $k = 1, \ldots, K$, like in Wei et al. (2012), we take the median of each profile, denoted by δ_k, as its center. Next we consider the shape. The profile medians are subsequently subtracted from each profile, such that all profiles are brought to a comparable level. To take the censoring rate and the samples size into account, before continuing with the decomposition of shape and variation, we first normalize each profile by its standard error at each time point along the time grid. This gives

$$\widetilde{\Lambda}_k(t_j) = \frac{\widehat{\Lambda}_{n_k,k}(t_j) - \delta_k}{\mathrm{sd}(\widehat{\Lambda}_{n_k,k}(t_j))}, \quad j = 1, \ldots, G, \; k = 1, \ldots, K. \tag{4.7}$$

With the locations of $\widehat{\Lambda}_{n_k,k}$ accounted for, we consider the shape-scale model for $\widetilde{\Lambda}_k$:

$$\widetilde{\Lambda}_k(t_j) = \mu(t_j) + s(t_j)e_{k,j}, \quad j = 1, \ldots, G, \quad k = 1, \ldots K,$$

where $e_{k,j}$ comes from a stationary process such that $\mathrm{median}(e_{k,j}) = 0$ and $\mathrm{median}(|e_{k,j}|) = 1$. Denoting the radial basis function (RBF) kernel with bandwidth b as $W_b(\cdot)$, the least absolute deviation (LAD) estimate of the shape function $\widehat{\mu}_b(t)$ characterizing the median of the normalized curves at each time t_j can be obtained by numerically solving the following minimization problem:

$$\widehat{\mu}_b(t) = \arg\min_{\theta} \sum_{k=1}^{K} \sum_{j=1}^{G} \left| \widetilde{\Lambda}_k(t_j) - \theta \right| W_b(t_j - t).$$

Noting that $\widehat{\mu}_b(t)$ could be biased, Wei et al. (2012) also gave the bias-corrected jackknife estimator,

$$\widetilde{\mu}_b(t) = 2\widehat{\mu}_b(t) - \widehat{\mu}_{\sqrt{2}b}(t).$$

Next, the reference deviation function $\widehat{s}_h(t)$ is estimated using the same LAD approach based on the residuals from the previous step:

$$\widehat{s}_h(t) = \arg\min_{\theta} \sum_{k=1}^{K} \sum_{j=1}^{G} \left| |\widetilde{\Lambda}_k(t_j) - \widetilde{\mu}_b(t)| - \theta \right| W_h(t_j - t),$$

where h is another bandwidth for the RBF kernel. The bias-corrected jackknife estimator of $s(t)$ is obtained as

$$\widetilde{s}_h(t) = 2\widehat{s}_h(t) - \widehat{s}_{\sqrt{2}h}(t).$$

The optimal values of b and h can be obtained using leave-one-out cross validation on the Phase I profiles.

Based on the estimated reference shape and variation, three descriptive statistics can be calculated, which describe the reasonable range of variabilities within the Phase I profiles. For vertical shift of the profile centers, we have

$$D_k = \frac{|\delta_k - \widehat{\mu}_\delta|}{\widehat{s}_\delta}, \tag{4.8}$$

where $\widehat{\mu}_\delta$ is the sample median of the Phase I profile centers $(\delta_1, \ldots, \delta_K)$, and $\widehat{s}_\delta$ is the their median absolute deviation. Next, based on the censoring rate and sample size adjusted curves $\widetilde{\Lambda}_k(t)$, we can obtain a curve of normalized deviations as

$$\widetilde{e}_k(t_j) = \left\{\widetilde{\Lambda}_k(t_j) - \widetilde{\mu}_b(t_j)\right\} / \widetilde{s}_h(t_j), \quad j = 1, \ldots, G, \tag{4.9}$$

and obtain two measures of shape deviation as:

$$T_{k,1} = \max_j |\widetilde{e}_k(t_j)|, \quad T_{k,2} = \sum_{j=1}^{G} |\widetilde{e}_k(t_j)|,$$

with $T_{k,1}$ being the maximum absolute shape deviation, and $T_{k,2}$ being the cumulative absolute shape deviation.

For a Phase II cumulative baseline hazard profile $\widehat{\Lambda}^*_{n_\ell,\ell}(t)$, we first record its center and denote it as δ^*_ℓ, and then calculate the normalized statistic like in (4.8):

$$D^*_\ell = \frac{|\delta^*_\ell - \widehat{\mu}_\delta|}{\widehat{s}_\delta}.$$

It is then adjusted for sample size and censoring rate, and we obtain

$$\widetilde{\Lambda}^*_\ell(t_j) = \frac{\widehat{\Lambda}^*_{n_\ell,\ell}(t_j) - \delta^*_\ell}{\mathrm{sd}(\widehat{\Lambda}^*_{n_\ell,\ell}(t_j))}, \quad j = 1, \ldots, G. \tag{4.10}$$

Based on (4.10), we can again obtain the curve of normalized deviations as

$$\widetilde{e}^*_\ell(t_j) = \{\widetilde{\Lambda}^*_\ell(t_j) - \widetilde{\mu}_b(t_j)\}/\widetilde{s}_h(t_j), \quad j = 1, \ldots, G,$$

and the two measures of shape deviation of the ℓth Phase II profile from the reference estimated from Phase I:

$$T^*_{\ell,1} = \max_j |\widetilde{e}^*_\ell(t_j)|, \quad T^*_{\ell,2} = \sum_{j=1}^{G} |\widetilde{e}^*_\ell(t_j)|, \quad \ell = 1, \ldots. \tag{4.11}$$

To decide whether a Phase II profile is "out of control" or not, we need to set the critical values of the three summary statistics. Let $c^{(0)}(\alpha)$, $c^{(1)}(\alpha)$, and $c^{(2)}(\alpha)$ be the $100(1-\alpha)$th percentiles of the empirical distributions for D_k, $T_{k,1}$ and $T_{k,2}$. For a given significance level α_0, Wei et al. (2012) proposed a numerical approach to determine the critical values by setting α to be the α^* such that

$$\alpha^* = \max_\alpha \left\{ \alpha : \sum_{k=1}^{K} \max\{1_{D_k > c^{(0)}(\alpha)}, 1_{T_{k,1} > c^{(1)}(\alpha)}, 1_{T_{k,2} > c^{(2)}(\alpha)}\} \le K\alpha_0 \right\}. \tag{4.12}$$

The critical values are then $c^{(0)}(\alpha^*)$, $c^{(1)}(\alpha^*)$, and $c^{(2)}(\alpha^*)$, respectively, for statistics D^*_ℓ,

$T^*_{\ell,1}$ and $T^*_{\ell,2}$.

4.3.3 Simultaneous Monitoring

For the Cox model, we extend (4.12) by incorporating the Hotelling's T^2 statistic. We now have four summary statistics, one corresponding to the parametric component and the other three corresponding to the nonparametric component. For a given desired significance level α_0, choose α^* such that

$$\alpha^* = \max_{\alpha} \left\{ \alpha : \sum_{k=1}^{K} \max\{1_{D_k > c^{(0)}(\alpha)}, 1_{T_{k,1} > c^{(1)}(\alpha)}, 1_{T_{k,2} > c^{(2)}(\alpha)}, 1_{\widetilde{T}^2_k > c^{(3)}(\alpha)}\} \le K\alpha_0 \right\}, \tag{4.13}$$

where $c^{(3)}(\alpha)$ is the $100(1-\alpha)$th percentile of the empirical distribution of $\widetilde{T}^2_k$. A Phase II block will be considered "out of control" if either one of the four statistics exceeds its the critical values $c^{(0)}(\alpha^*)$, $c^{(1)}(\alpha^*)$, $c^{(2)}(\alpha^*)$, and $c^{(3)}(\alpha^*)$. The approach in (4.13) ensures that empirically, among the K Phase I blocks, no more than $K\alpha_0$ are identified to be "out of control". The empirical rejection rate of the procedure on Phase II blocks is

$$\frac{1}{L} \sum_{\ell=1}^{L} \max\{1_{D^*_\ell > c^{(0)}(\alpha^*)}, 1_{T^*_{\ell,1} > c^{(1)}(\alpha^*)}, 1_{T^*_{\ell,2} > c^{(2)}(\alpha^*)}, 1_{\widetilde{T}^{2*}_\ell > c^{(3)}(\alpha^*)}\},$$

where L is the total number of Phase II blocks.

4.4 Simulation Studies

4.4.1 Size

We simulate normative Phase I blocks of data from a Cox proportional hazards model as the base to construct the proposed test statistics and their control limits. For the linear predictor of Cox model, we choose $\beta = (0.7, -0.5, 0.4)^\top$ that corresponds to three covariates: $x_{ki[1]} \sim N(0,1)$, $x_{ki[2]} \sim$ Bernoulli(0.5), and $x_{ki[3]} \sim$ Bernoulli(0.1) independently. Next, we specify two baseline hazard functions, $\lambda_{01}(t) = 0.02$ and $\lambda_{02}(t) = 0.06t^{0.7}$, corresponding to, respectively, the exponential and Weibull distributions if there were no covariates. To account for different censoring rates in blocks, for block k, $k = 1, \ldots, K$, the censoring times are generated from a mixture distribution $\epsilon_k \langle 60 \rangle + (1 - \epsilon_k)\text{Unif}(0, 60)$ where $\langle 60 \rangle$ stands for a point mass at 60, and $\epsilon_k \sim \text{Unif}(0.1, 0.9)$. The censoring rates range from approximately 40% when $\epsilon = 0.9$, and 60% when $\epsilon = 0.1$. For the exponential baseline hazard scenario, the bandwidths are cross-validated on 500 Phase I profiles and $b_{\text{Exp}} = 0.3$, $h_{\text{Exp}} = 0.8$ are used. For the Weibull case, $b_{\text{Weib}} = 0.7$, and $h_{\text{Weib}} = 0.3$.

We considered cases where there are $K \in \{200, 400\}$ normative Phase I blocks, respectively. To assess the performance of the proposed screening procedure for differences in block sizes, the block sizes are generated using a uniform distribution over the integers from 2000 to 4000. In each replicate of simulation, after the K normative blocks are generated, the four statistics are calculated, and the procedure in (4.13) is used to choose the α^* corresponding to an overall $\alpha_0 = 0.05$. Then 500 Phase II blocks are simulated using the same setting, and the four statistics for these blocks are again calculated and compared with their respective control limits. The proportion of Phase II blocks classified as "violators" is calculated, which is the empirical Type I error. A total of 500 replicates are run for each combination of K and n_k. The average Type I errors are reported in Table 5. The proposed screening procedure has size close the nominal level of 0.05 for all combinations of block sizes and numbers of Phase I blocks. For even bigger K, such as 600 or 800, the empirical Type I error rates are even closer to 0.05. In addition, we recorded the proportions of rejections made by each statistic, and the detailed results are included in Appendix A.

4.4.2 Power

To demonstrate the power of the proposed monitoring scheme, we generated Phase II blocks from different alternative models where either the baseline hazard or the vector of coefficients departs from the model from which Phase I data is generated. Then the proportion of blocks identified as "violators" is calculated, which is the empirical power of the proposed test. To study whether the "correct" statistic picked up the violation, we also recorded the individual powers of the four statistics. The details are given in Appendix A.

The baseline hazard for Phase II blocks can vary in multiple ways. For different blocks, the baseline hazard may fluctuate around a certain value, but with some variance; or, the baseline hazard can be greater or smaller than the Phase I blocks in an overall manner. For the increased variability scenario, all Phase I blocks are generated using the same setting as in Section 4.4.1. Phase II blocks are generated using the same setting, except that the the hazard rate becomes

$$\begin{aligned}\lambda'_{01_i} &= 0.02 + \epsilon_i \\ \lambda'_{02_i} &= 0.06 t^{0.7+\varepsilon_i}\end{aligned}$$

with $\epsilon_i \sim \mathrm{Unif}(-0.005, 0.005)$, $\varepsilon_i \sim \mathrm{Unif}(-0.1, 0.1)$ corresponding to small variability, and $\epsilon_i \sim \mathrm{Unif}(-0.01, 0.01)$, $\varepsilon_i \sim \mathrm{Unif}(-0.2, 0.2)$ corresponding to large variability. We considered two overall shifts as well. Phase II blocks are generated using the same setting, except that the hazard rate becomes

$$\begin{aligned}\lambda'_{01_i} &= 0.02 + \delta_{v1}, \\ \lambda'_{02_i} &= (0.06 + \delta_{v_2}) t^{0.7},\end{aligned}$$

with $\delta_{v_1} = 0.0015$, $\delta_{v_2} = 0.005$ for a small shift, and $\delta_{v_1} = 0.003$, $\delta_{v_2} = 0.01$ for a large shift. The obtained average empirical powers on 500 replicates are reported in Table 5. For both baseline hazards, the power increases with the magnitude of change. For different K's, the fluctuation in power is rather small, indicating that for the scenario presented here, 200 Phase I blocks provide enough information for the proposed procedure to identify violators.

Table 5: Average Size and Power under different changes for Phase II data.

Component	Change	λ_{01}		λ_{02}	
		$K = 200$	$K = 400$	$K = 200$	$K = 400$
No Change, under the null hypothesis		0.062	0.058	0.062	0.057
Baseline, λ_0	Increased Variability				
	small increase	0.601	0.600	0.724	0.720
	large increase	0.800	0.800	0.862	0.860
	Overall Shift				
	small shift	0.294	0.286	0.376	0.368
	large shift	0.846	0.846	0.924	0.926
Parameter, β	Increased Variability				
	$\Sigma_\epsilon = 0.04I_{3\times3}$	0.263	0.258	0.271	0.270
	$\Sigma_\epsilon = 0.08I_{3\times3}$	0.629	0.627	0.638	0.637
	Shift in Covariate Effect				
	$\Delta\beta_2 = -0.15$	0.473	0.468	0.741	0.741
	$\Delta\beta_2 = -0.3$	0.983	0.984	0.957	0.957

The parametric part can also change in multiple ways. For different blocks, like the baseline hazard, the vector of coefficients may fluctuate around a certain vector, but with some variance; or, the effect of one particular covariate can change. Phase I blocks are generated using the same setting as before. For Phase II blocks, the vector of coefficients is obtained as $\beta_{\text{Phase II}} = (0.7, -0.5, 0.4) + \varepsilon$, with $\varepsilon \sim N(0, 0.04I_{3\times3})$ corresponding to small variability, and $\varepsilon \sim N(0, 0.08I_{3\times3})$ corresponding to large variability. For the case where a single covariate effect changes, we considered two alternative β's: $(0.7, -0.65, 0.4)$, and $(0.7, -0.8, 0.4)$. A total of 500 replicates are run for each scenario specified above. The obtained average empirical powers are also reported in Table 5.

Again, as expected, the power increases as the variability of parameter vector or magnitude of shift in covariate effect increases. For both the exponential and Weibull baselines, when variability of parameter vector increases, the powers are similar. When covariate effect shifts, the power is higher for the Weibull baseline case. Similar patterns as in Table 5 for changes in λ_0 are observed. The power increases when the magnitude of change increases. The influence of K is again insignificant here. This indicates that, 200 Phase I blocks is enough to establish credible control limits for the four statistics.

4.5 Monitoring the Survival of Lymphoma Patients

We choose a subset of the SEER lymphoma data diagnosed between 1974 and 1998 for illustration. The selected dataset consists of 84,794 patients, out of which 33,557 had events due to lymphoma. The same 3-month partition scheme was used, resulting in 100 blocks of survival data.

As a starting point, we partitioned the data as in Qiu et al. (2010) such that 2/3 of all blocks serve as Phase I, i.e., blocks 1 to 66 are used as Phase I, and blocks 67 to 100 are used as Phase II. This is also consistent with the findings in Chapter 3, where the proposed online updating cumulative test statistic suggested a significant change in the underlying model around approximately the same time. A Cox regression is run with three covariates: a continuous covariate Age, and two indicators Female and Black, and $\widehat{\beta}_{n_k,k}$ and $\mathcal{I}_{n_k,k}(\widehat{\beta}_{n_k,k})$ for $k = 1, \ldots, 66$ are obtained. We subsequently calculated $\widetilde{\beta}_1, \ldots, \widetilde{\beta}_{66}$ and $\overline{\widetilde{\beta}}_{66}$, and obtained the $\widetilde{T}^2$ statistics calculated using (4.5). For the nonparametric component, as in SEER reporting, survival times are reported in integer months and censored at 60 months, a natural choice for the time grid is the set of integer values from 1 to 60. The Breslow estimators $\widehat{\Lambda}_{n_1,1}, \ldots, \widehat{\Lambda}_{n_{66},66}$, are obtained, each consisting of 60 (time, cumulative hazard) pairs. Due to the heterogeneity in block sizes and censoring rates, after centering each curve with its median δ_k, we used the normalization in (4.7) and obtained $\widetilde{\Lambda}_k$ for $k = 1, \ldots, 66$. We then estimate the reference curve and variability as in Wei et al. (2012). Again, leave-one-out cross validation was used to select the optimal bandwidths among $\{0.1, 0.2, \ldots, 3\}$ for the RBF kernels for, respectively, the shape $\widetilde{\mu}(t)$, and the variability $\widetilde{s}(t)$, at times $t = 1, \ldots, 60$. The final bandwidth b for shape was selected to be 0.9, and h for variability was selected to be 1.7. Based on the selected bandwidths, the relative vertical deviations D_k, the maximum absolute shape deviations $T_{k,1}$, and cumulative absolute shape deviations $T_{k,2}$ are calculated. The α^* for each individual statistic was selected to be 0.01535 so that they jointly identified 4 violators within the Phase I blocks, yielding an overall Phase I α_0 of 0.061. Their control limits are subsequently obtained by taking the $100(1-\alpha^*)$th percentile of their empirical distributions.

For patients in the rest blocks, the Phase II statistics, $\widetilde{T}_\ell^{2*}$, D_ℓ^*, $T_{\ell,1}^*$, and $T_{\ell,1}^*$ are calculated as in (4.6) and (4.11). The four statistics for Phase I and Phase II blocks are plotted together in Figure 23. The control limits established in Phase I are plotted as horizontal dashed lines. Phase I and Phase II statistics are separated using vertical

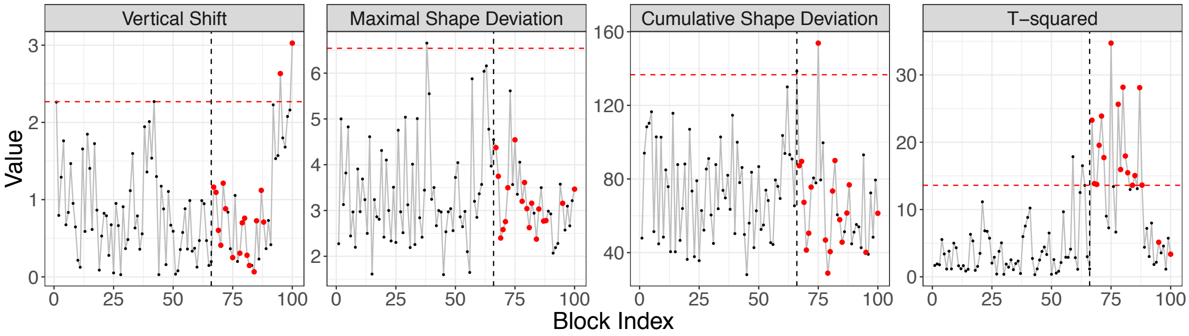

Figure 23: Profile charts of the four measurements for each block. Their respective thresholds are plotted using red dashed line. Outliers that they jointly identified are plotted in red.

dashed lines. It can be seen that within the 32 Phase II blocks, 11 have been identified as violators. Blocks 95 and 100 have been singled out due to relatively large vertical shifts. Block 75 has an abnormally large cumulative shape deviation compared to the reference obtained in Phase I. The T^2 statistics of blocks 67, 70, 71, 72, 75, 78, 80, 81 and 87 are too large to be normative.

To verify that this is indeed the case, we plot the estimated cumulative baseline hazard curves, in panel (a) of Figure 24, we plot the profiles obtained for Phase II blocks, where the two dashed lines are the violators identified by D^*_ℓ. In panel (b), the absolute normalized shape deviations ($|(\widetilde{\Lambda}^*_\ell(t) - \widetilde{\mu}(t))|/\widetilde{s}(t)$) are plotted, where the dashed line corresponds to block 75. In panel (c), boxplots for normalized Phase I parameter estimates $\widetilde{\beta}_k$ for $k = 1, \ldots, 66$ are first drawn, and normalized Phase II parameter estimates ($\widetilde{\beta}^*_\ell$) are overlaid using triangles to differentiate from one outlier in Phase I, which was plotted as a dot. It is obvious that the covariate effects are outlying when compared to Phase I.

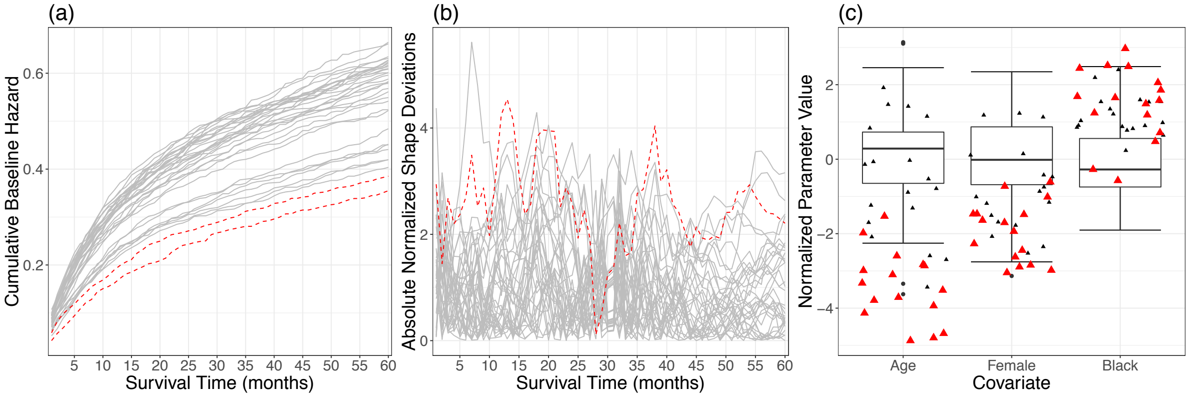

Figure 24: (a) The original curves $\widehat{\Lambda}_{n_{67},67}, \ldots, \widehat{\Lambda}_{n_{100},100}$; (b) the absolute normalized shape deviations $(|(\widetilde{\Lambda}^*_\ell(t) - \widetilde{\mu}(t))|/\widetilde{s}(t))$ for $\ell = 67, \ldots, 100$; (c) boxplot of Phase I normalized parameters, with Phase II normalized parameters scattered as triangles.

4.6 Discussion

While the online updating approach help us decide whether there is an violation in the parametric component of the Cox model along the data stream, a tool is also needed to detect changes in the baseline hazards, which was the motivation for this chapter. We proposed a procedure that is based on a collection of Cox model estimates obtain on known, “in control” Phase I blocks of survival data, and detects departures of Phase II data blocks from Phase I, in terms of both the parametric component β, and the nonparametric baseline hazard. This approach was inspired by the profile monitoring methods of Williams et al. (2007) for regression coefficients and of Wei et al. (2012) for curves. The approach in Wei et al. (2012) is extended to incorporate four test statistics to ensure that, when Phase II data are in control, the empirical type I error rate is appropriately controlled. One novelty of the proposed method is that it allows different sample sizes in different blocks through a proper normalization procedure, which ensures that the determination of control limits is not impacted by block sizes. Block sizes, however, should in general be sufficiently large, so that the asymptotic distributions of $\widetilde{\beta}_k$ and $\widetilde{\Lambda}_k$ are valid. In addition, the validity of the proposed approach resides on the assumption that within each block, the observations are generated from the same underlying model. As the control limits are established using quantiles of the empirical distribution of Phase I statistics, and are essentially approximations of the true underlying quantiles, it is reasonable to expect that as the number of Phase I blocks increase, the more precise the approximation becomes.

The choice of time grid for the Breslow estimators should not be arbitrary. On the one hand, we want to choose enough time points, so that the obtained profile is representative of the Breslow estimator curve. On the other hand, when the grid is chosen to be too fine, the corresponding computing burden would be heavy. Neither will the approximation be necessary when the number of time points exceeds the number of event times in each block.

In simulation studies, the proposed simultaneous monitoring method holds its size when Phase II blocks are generated from the same model as Phase I blocks, and the size approaches its nominal level of 0.05 with the increase in number of Phase I profiles, while having substantial power in detecting different types of model change, including shifts in, or increase in variabilities of, either the baseline hazard rate, or the covariate effects. In the application to SEER lymphoma data, the proposed procedure identified 11 outlying blocks in Phase II.

The proposed method can be extended to the monitoring of other semi-parametric models, such as the proportional odds model (Bennett, 1983) and the additive hazards model (Cox and Oakes, 1984), and out of the survival analysis context, the partially linear model (Robinson, 1988) and index model (Ichimura, 1993), provided that appropriate approaches bring both components to a comparable level.

Chapter 5

Future Work

In Chapter 4, when monitoring both the parametric and nonparametric components of the Cox model, we are using three nonparametric statistics designed for depicting curves, and we are interested in whether Phase II blocks are different or not, without regards to any increasing or decreasing trend. If our alternative hypothesis is not “a Phase II block is outlying”, but “Phase II blocks are having overall higher cumulative baseline hazard”, testing methods based on contrasts similar to those in Hu and Huffer (2019), where the Nelson–Aalen estimator and Kaplan–Meier estimators are studied, can be constructed for the Breslow estimators. In addition, when data keep arriving, gradual changes may happen. For such cases, using fixed control limits for statistics might raise too many false rejections. How to dynamically and adaptively set the control limits to accommodate the gradual changes would be an interesting topic.

For a rather general online updating setting, variable selection remains an interesting topic. It is online updating’s natural advantage that previous parameter estimates and variable selection results can be used to inform selection of weights in fitting an adaptive Lasso (Zou, 2006) procedure for the current block. Under the divide and conquer setting, Chen and Xie (2014) used a majority voting approach to select the final set of covariates. Tang et al. (2016) constructed a confidence distribution that allows combining results from multiple blockwise analysis results. Development of similar or more precise approaches under the online updating setting is devoted to future research.

Appendix A

Simultaneous Monitoring Supplementation

A.1 Detailed Simulation Results

In addition to the results presented in Table 5 in the main text, we also recorded which of the four statistics exceeded their respective thresholds in each simulation scenario.

It can be seen that the overall size is very close to the nominal level of 0.05, and the proportions of violators identified by each of the four statistics are very close.

When the variability of the baseline hazard function increases in Phase II, as indicated in Table 7, in addition to D_i, the two statistics describing shape deviation also identified an increased proportion of outliers. The effect is more obvious for the Weibull distribution, as changing the Weibull scale parameter will give more complicated variations in the shape of the cumulative baseline hazard curve than a simple shift. When we tweak the baseline hazard function by changing the value of a parameter, from Table 8, while the overall power demonstrates the effectiveness of the combined rejection rule, it

Table 6: Average proportions of rejections given by each statistic under the null hypothesis

Baseline	K	Overall Size	D_i	$T^*_{i,1}$	$T^*_{i,2}$	$\widetilde{T}^2_i$
Exponential	200	0.062	0.017	0.018	0.017	0.017
	400	0.058	0.016	0.017	0.017	0.015
Weibull	200	0.062	0.017	0.019	0.017	0.017
	400	0.057	0.016	0.017	0.016	0.016

Table 7: Average proportions of rejections given by each statistic when the baseline hazard has an increase in variability

	Baseline	K	Overall Power	D_i	$T^*_{i,1}$	$T^*_{i,2}$	$\widetilde{T}^2_i$
Small Increase	Exponential	200	0.601	0.580	0.027	0.035	0.017
		400	0.600	0.581	0.024	0.033	0.016
	Weibull	200	0.724	0.701	0.181	0.198	0.028
		400	0.720	0.698	0.179	0.198	0.027
Large Increase	Exponential	200	0.800	0.789	0.076	0.093	0.017
		400	0.800	0.790	0.070	0.092	0.016
	Weibull	200	0.862	0.851	0.524	0.507	0.036
		400	0.860	0.849	0.524	0.505	0.035

can be found that it's the statistic for location shift, D_i, that plays a major role here in identifying violators.

For the vector of coefficients, as it is relatively independent from the first three statistics that describe the cumulative baseline hazard profile, in Tables 9 and 10, when comparing the results to Table 6, it can be found that the proportions of violators identified by the first three statististics had little change. It is $\widetilde{T}^2_i$ that made the major contribution in identifying outliers.

Table 8: Average proportions of rejections given by each statistic when the baseline hazard has a shift

	Baseline	K	Overall Power	D_i	$T^*_{i,1}$	$T^*_{i,2}$	$\widetilde{T}^2_i$
Small Shift	Exponential	200	0.294	0.254	0.024	0.024	0.016
		400	0.286	0.248	0.022	0.023	0.015
	Weibull	200	0.376	0.334	0.025	0.027	0.023
		400	0.368	0.330	0.023	0.025	0.021
Large Shift	Exponential	200	0.846	0.834	0.034	0.040	0.017
		400	0.846	0.835	0.031	0.039	0.015
	Weibull	200	0.924	0.917	0.035	0.044	0.019
		400	0.926	0.920	0.032	0.043	0.019

Table 9: Average proportions of rejections given by each statistic when the coefficient vector has an increase in variability

	Baseline	K	Overall Power	D_i	$T^*_{i,1}$	$T^*_{i,2}$	$\widetilde{T}^2_i$
Small Increase	Exponential	200	0.263	0.017	0.018	0.017	0.229
		400	0.258	0.015	0.016	0.015	0.227
	Weibull	200	0.271	0.017	0.019	0.017	0.238
		400	0.270	0.016	0.017	0.016	0.239
Large Increase	Exponential	200	0.629	0.017	0.019	0.017	0.611
		400	0.627	0.016	0.017	0.016	0.610
	Weibull	200	0.638	0.017	0.019	0.018	0.621
		400	0.637	0.016	0.018	0.017	0.621

Table 10: Average proportions of rejections given by each statistic when the coefficient vector has a shift

	Baseline	K	Overall Power	D_i	$T^*_{i,1}$	$T^*_{i,2}$	$\widetilde{T}^2_i$
Small Shift	Exponential	200	0.473	0.017	0.019	0.017	0.448
		400	0.468	0.015	0.017	0.016	0.446
	Weibull	200	0.741	0.018	0.019	0.018	0.730
		400	0.741	0.016	0.018	0.016	0.730
Large Shift	Exponential	200	0.983	0.018	0.021	0.018	0.979
		400	0.984	0.016	0.018	0.017	0.980
	Weibull	200	0.957	0.019	0.022	0.020	0.955
		400	0.957	0.017	0.020	0.019	0.955

Bibliography

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.

Andersen, P. K. (1982). Testing goodness of fit of Cox's regression and life model. *Biometrics 38*(1), 67–77.

Andersen, P. K. and R. D. Gill (1982). Cox's regression model for counting processes: a large sample study. *The Annals of Statistics 10*(4), 1100–1120.

Arjas, E. (1988). A graphical method for assessing goodness of fit in Cox's proportional hazards model. *Journal of the American Statistical Association 83*(401), 204–212.

Barlow, W. E. (1997). Global measures of local influence for proportional hazards regression models. *Biometrics 53*(3), 1157–1162.

Barlow, W. E. and R. L. Prentice (1988). Residuals for relative risk regression. *Biometrika 75*(1), 65–74.

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

Bennett, S. (1983). Analysis of survival data by the proportional odds model. *Statistics in Medicine 2*(2), 273–277.

Breslow, N. (1972). Discussion of the paper "Regression Models and Life-Tables" by Dr Cox. *Journal of the Royal Statistical Society. Series B (Methodological) 34*(2), 216–217.

Cai, Z. and Y. Sun (2003). Local linear estimation for time-dependent coefficients in Cox's regression models. *Scandinavian Journal of Statistics 30*(1), 93–111.

Cain, K. C. and N. T. Lange (1984). Approximate case influence for the proportional hazards regression model with censored data. *Biometrics 40*(2), 493–499.

Caplan, D., Y. Li, W. Wang, S. Kang, L. Marchini, H. Cowen, and J. Yan (2019). Dental restoration longevity among geriatric and special needs patients. *JDR Clinical & Translational Research 4*(1), 41–48.

Chang, W. and J. Luraschi (2018). *profvis: Interactive Visualizations for Profiling R Code*. R package version 0.3.5.

Chappell, R. (1992). A note on linear rank tests and Gill and Schumacher's tests of proportionality. *Biometrika 79*(1), 199–201.

Chen, K., H. Lin, and Y. Zhou (2012). Efficient estimation for the Cox model with varying coefficients. *Biometrika 99*(2), 379–392.

Chen, X. and M. Xie (2014). A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica 24*(4), 1655–1684.

Collett, D. (2015). *Modelling Survival Data in Medical Research*. CRC press.

Cook, R. D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society. Series B (Methodological) 48*(2), 133–169.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological) 34*(2), 187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika 62*(2), 269–276.

Cox, D. R. (1979). A note on the graphical analysis of survival data. *Biometrika 66*(1), 188–190.

Cox, D. R. and D. Oakes (1984). *Analysis of Survival Data*, Volume 21. CRC Press.

Cox, D. R. and E. J. Snell (1968). A general definition of residuals. *Journal of the Royal Statistical Society. Series B (Methodological) 30*(2), 248–275.

Crowley, J. and M. Hu (1977). Covariance analysis of heart transplant survival data. *Journal of the American Statistical Association 72*(357), 27–36.

Fan, J., Y. Feng, and Y. Wu (2010). High-dimensional variable selection for Cox's proportional hazards model. In *Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown*, pp. 70–86. Institute of Mathematical Statistics.

Fan, J. and R. Li (2002). Variable selection for Cox's proportional hazards model and frailty model. *Annals of Statistics 30*(1), 74–99.

Fan, J., H. Lin, and Y. Zhou (2006). Local partial-likelihood estimation for lifetime data. *The Annals of Statistics 34*(1), 290–325.

Farrington, C. P. (2000). Residuals for proportional hazards models with interval-censored survival data. *Biometrics 56*(2), 473–482.

Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics 42*(4), 845–854.

Fisher, L. D., , and D. Y. Lin (1999). Time-dependent covariates in the Cox proportional-hazards regression model. *Annual Review of Public Health 20*(1), 145–157.

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

Gill, R. and M. Schumacher (1987). A simple test of the proportional hazards assumption. *Biometrika 74*(2), 289–300.

Goeman, J. J. (2009). L1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal 52*(1), 70–84.

Goggins, W. B. and D. M. Finkelstein (2004). A proportional hazards model for multivariate interval-censored failure time data. *Biometrics 56*(3), 940–943.

Grambsch, P. M. (1995). Goodness-of-fit and diagnostics for proportional hazards regression models. In *Recent Advances in Clinical Trial Design and Analysis*, pp. 95–112. Springer.

Grambsch, P. M. and T. M. Therneau (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika 81*(3), 515–526.

Grant, S., Y. Q. Chen, and S. May (2014). Performance of goodness-of-fit tests for the Cox proportional hazards model with time-varying covariates. *Lifetime Data Analysis 20*(3), 355–368.

Grønnesby, J. K. and Ø. Borgan (1996). A method for checking regression models in survival analysis based on the risk score. *Lifetime Data Analysis 2*(4), 315–328.

Harrell, F. E. (1986). The PHGLM Procedure. In *SUGI Supplemental Library Users Guide* (Version 5 ed.)., pp. 437–466. Cary, NC: SAS Institute, Inc.

Hastie, T. and R. Tibshirani (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological) 55*(4), 757–796.

Heller, G. (2010). Proportional hazards regression with interval censored data using an inverse probability weight. *Lifetime Data Analysis 17*(3), 373–385.

Henderson, R. and A. Milner (1991). On residual plots for relative risk regression. *Biometrika 78*(3), 631–636.

Hess, K. R. (1995). Graphical methods for assessing violations of the proportional hazards assumption in Cox regression. *Statistics in Medicine 14*(15), 1707–1723.

Holmes, D. S. and A. E. Mergen (1993). Improving the performance of the T^2 control chart. *Quality Engineering 5*(4), 619–625.

Hotelling, H. (1931). The generalization of student's ratio. *The Annals of Mathematical Statistics 2*(3), 360–378.

Hu, G. and F. Huffer (2019). Modified Kaplan–Meier estimator and Nelson–Aalen estimator with geographical weighting for survival data. *Geographical Analysis*. Forthcoming.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics 58*(1–2), 71–120.

Javanmard, A. and A. Montanari (2018). Online rules for control of false discovery rate and false discovery exceedance. *The Annals of Statistics 46*(2), 526–554.

Jensen, W. A., J. B. Birch, and W. H. Woodall (2007). High breakdown estimation methods for phase I multivariate control charts. *Quality and Reliability Engineering International 23*(5), 615–629.

Jones, B. S. and R. P. Branton (2005). Beyond logit and probit: Cox duration models of single, repeating, and competing events for state policy adoption. *State Politics & Policy Quarterly 5*(4), 420–443.

Kalbfleisch, J. D. and R. L. Prentice (2002). *The Statistical Analysis of Failure Time Data*. New York; Chichester: John Wiley & Sons.

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

Kang, L. and S. L. Albin (2000). On-line monitoring when the process yields a linear profile. *Journal of Quality Technology 32*(4), 418–426.

Kassambara, A. and M. Kosinski (2017). *survminer: Drawing Survival Curves using "ggplot2"*. R package version 0.4.0.

Kay, R. (1977). Proportional hazard regression models and the analysis of censored survival data. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 26*(3), 227–237.

Kazemzadeh, R. B., R. Noorossana, and A. Amiri (2008). Phase I monitoring of polynomial profiles. *Communications in Statistics – Theory and Methods 37*(10), 1671–1686.

Keele, L. (2010). Proportionally difficult: Testing for nonproportional hazards in Cox models. *Political Analysis 18*(2), 189–205.

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

Kotz, S. and N. Johnson (1992). *Breakthrough in Statistics: Volume I, Foundations and Basic Theory.* Berlin; New York: Springer–Verlag Inc.

Kupets, O. (2006). Determinants of unemployment duration in Ukraine. *Journal of Comparative Economics 34*(2), 228–247.

Lagakos, S. (1981). The graphical evaluation of explanatory variables in proportional hazard regression models. *Biometrika 68*(1), 93–98.

Lane, W. R., S. W. Looney, and J. W. Wansley (1986). An application of the Cox proportional hazards model to bank failure. *Journal of Banking & Finance 10*(4), 511–531.

Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data.* New York; Chichester: John Wiley & Sons.

Lawrance, A. J. (1995). Deletion influence and masking in regression. *Journal of the Royal Statistical Society. Series B (Methodological) 57*(1), 181–189.

Lee, S.-H. and C.-H. Jun (2010). A new control scheme always better than X-bar chart. *Communications in Statistics – Theory and Methods 39*(19), 3492–3503.

Lee, S.-H. and C.-H. Jun (2012). A process monitoring scheme controlling false discovery rate. *Communications in Statistics – Simulation and Computation 41*(10), 1912–1920.

Lin, D. (2007). On the Breslow estimator. *Lifetime Data Analysis 13*(4), 471–480.

Lin, D. Y. (1991). Goodness-of-fit analysis for the Cox regression model based on a class of parameter estimators. *Journal of the American Statistical Association 86*(415), 725–728.

Lin, D. Y. and L. J. Wei (1989). The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association 84*(408), 1074–1078.

Lin, D. Y., L. J. Wei, and Z. Ying (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika 80*(3), 557–572.

Lin, N. and R. Xi (2011). Aggregated estimating equation estimation. *Statistics and Its Interface 4*(1), 73–83.

Marzec, L. and P. Marzec (1997a). Generalized martingale-residual processes for goodness-of-fit inference in Cox's type regression models. *The Annals of Statistics 25*(2), 683–714.

Marzec, L. and P. Marzec (1997b). On fitting Cox's regression model with time-dependent coefficients. *Biometrika 84*(4), 901–908.

McCullagh, P. and J. A. Nelder (1983). *Generalized Linear Models*. London; New York: Chapman & Hall Ltd.

Mittal, S., D. Madigan, R. S. Burd, and M. A. Suchard (2014). High-dimensional, massive sample-size Cox proportional hazards regression for survival analysis. *Biostatistics 15*(2), 207–221.

Moreau, T., J. O'Quigley, and J. Lellouch (1986). On D. Schoenfeld's approach for testing the proportional hazards assumption. *Biometrika 73*(2), 513–515.

Moreau, T., J. O'Quigley, and M. Mesbah (1985). A global goodness-of-fit statistic for the proportional hazards model. *Applied Statistics 34*(3), 212–218.

Murphy, S. A. and P. K. Sen (1991). Time-dependent coefficients in a Cox-type regression model. *Stochastic Processes and Their Applications 39*(1), 153–180.

Nagelkerke, N. J. D., J. Oosting, and A. A. M. Hart (1984). A simple test for goodness-of-fit of Cox's proportional hazards model. *Biometrics 40*, 483–486.

Nardi, A. and M. Schemper (1999). New residuals for Cox regression and ttheir application to outlier screening. *Biometrics 55*(2), 523–529.

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.

O'Quigley, J. and F. Pessione (1989). Score tests for homogeneity of regression effect in the proportional hazards model. *Biometrics 45*, 135–144.

Park, M. Y. and T. Hastie (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69*(4), 659–677.

Park, M. Y. and T. Hastie (2018). *glmpath: L1 Regularization Path for Generalized Linear Models and Cox Proportional Hazards Model*. R package version 0.98.

Park, S. and D. J. Hendry (2015). Reassessing Schoenfeld residual tests of proportional hazards in political science event history analyses. *American Journal of Political Science 59*(4), 1072–1087.

Pettitt, A. N. and I. B. Daud (1989). Case-weighted measures of influence for proportional hazards regression. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 38*(1), 51–67.

Qiu, P., C. Zou, and Z. Wang (2010). Nonparametric profile monitoring by mixed effects modeling. *Technometrics 52*(3), 265–277.

Reid, N. and H. Crépeau (1985). Influence functions for proportional hazards regression. *Biometrika 72*(1), 1–9.

Robinson, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica 56*(4), 931–954.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association 79*(388), 871–880.

Sargent, D. J. (1997). A flexible approach to time-varying coefficients in the Cox regression setting. *Lifetime Data Analysis 3*(1), 13–25.

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.

Schoenfeld, D. (1980). Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Biometrika 67*(1), 145–153.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika 69*(1), 239–241.

Schoenfeld, D. A. (1983). Sample-size formula for the proportional-hazards regression model. *Biometrics 39*(2), 499–503.

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

Storer, B. E. and J. Crowley (1985). A diagnostic for Cox regression and general conditional likelihoods. *Journal of the American Statistical Association 80*(389), 139–147.

Sullivan, J. H. and W. H. Woodall (1996). A comparison of multivariate control charts for individual observations. *Journal of Quality Technology 28*(4), 398–408.

Tang, L., L. Zhou, and P. X. K. Song (2016). Method of divide-and-combine in regularised generalised linear models for big data. *arXiv e-prints*, arXiv:1611.06208.

Tang, Y., M. Horikoshi, and W. Li (2016). ggfortify: Unified interface to visualize statistical result of popular R packages. *The R Journal*.

Therneau, T., C. Crowson, and E. Atkinson (2017). 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 P. M. Grambsch (2000). *Modeling Survival Data: Extending the Cox Model.* Berlin; New York: Springer–Verlag Inc.

Therneau, T. M., P. M. Grambsch, and T. R. Fleming (1990). Martingale-based residuals for survival models. *Biometrika 77*(1), 147–160.

Tian, L., D. Zucker, and L. Wei (2005). On the Cox model with time-varying regression coefficients. *Journal of the American statistical Association 100*(469), 172–183.

Vargas, N. J. A. (2003). Robust estimation in multivariate control charts for individual observations. *Journal of Quality Technology 35*(4), 367–376.

Verweij, P. J. M. and H. C. van Houwelingen (1995). Time-dependent effects of fixed covariates in Cox regression. *Biometrics 51*(4), 1550–1556.

Wang, K. and F. Tsung (2005). Using profile monitoring techniques for a data-rich environment with huge sample size. *Quality and Reliability Engineering International 21*(7), 677–688.

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

Wang, Y., N. Palmer, Q. Di, J. Schwartz, I. Kohane, and T. Cai (2018). A fast divide-and-conquer sparse Cox regression. *ArXiv e-prints*.

Webb, G. I. and F. Petitjean (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, New York, NY, USA, pp. 1255–1264. ACM.

Wei, L. J. (1984). Testing goodness of fit for proportional hazards model with censored observations. *Journal of the American Statistical Association 79*(387), 649–652.

Wei, W. H. and M. R. Kosorok (2000). Masking unmasked in the proportional hazards model. *Biometrics 56*(4), 991–995.

Wei, Y., Z. Zhao, and D. K. J. Lin (2012). Profile control charts based on nonparametric L_1 regression methods. *The Annals of Applied Statistics 6*(1), 409–427.

Weissfeld, L. A. (1990). Influence diagnostics for the proportional hazards model. *Statistics & probability letters 10*(5), 411–417.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis.* Springer–Verlag New York.

Williams, J. D., W. H. Woodall, and J. B. Birch (2007). Statistical monitoring of nonlinear product and process quality profiles. *Quality and Reliability Engineering International 23*(8), 925–941.

Winnett, A. and P. Sasieni (2001). Miscellanea. a note on scaled Schoenfeld residuals for the proportional hazards model. *Biometrika 88*(2), 565–571.

Woodall, W. H. (2007). Current research on profile monitoring. *Production 17*(3), 420–425.

Xue, X., X. Xie, M. Gunter, T. E. Rohan, S. Wassertheil-Smoller, G. Y. Ho, D. Cirillo, H. Yu, and H. D. Strickler (2013). Testing the proportional hazards assumption in case-cohort analysis. *BMC Medical Research Methodology 13*(1), 88.

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

Xue, Y. and E. D. Schifano (2017). Diagnostics for the Cox model. *Communications for Statistical Applications and Methods 24*(6), 583–604.

Yang, Y. and H. Zou (2013). A cocktail algorithm for solving the elastic net penalized Cox's regression in high dimensions. *Statistics and Its Interface 6*(2), 167–173.

Yang, Y. and H. Zou (2017). *fastcox: Lasso and Elastic-Net Penalized Cox's Regression in High Dimensions Models using the Cocktail Algorithm*. R package version 1.1.3.

Yu, G., C. Zou, and Z. Wang (2012). Outlier detection in functional observations with applications to profile monitoring. *Technometrics 54*(3), 308–318.

Zhou, H. and A. B. Lawson (2008). EWMA smoothing and Bayesian spatial modeling for health surveillance. *Statistics in Medicine 27*(28), 5907–5928.

Zhu, H., J. G. Ibrahim, and M.-H. Chen (2015). Diagnostic measures for the Cox regression model with missing covariates. *Biometrika 102*(4), 907–923.

Zhu, J. and D. K. J. Lin (2009). Monitoring the slopes of linear profiles. *Quality Engineering 22*(1), 1–12.

Zou, C., F. Tsung, and Z. Wang (2008). Monitoring profiles based on nonparametric regression methods. *Technometrics 50*(4), 512–526.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association 101*(476), 1418–1429.