Contemporary Clinical Trials Communications 8 (2017) 147–155

Contents lists available at ScienceDirect

Contemporary Clinical Trials Communications

journal homepage: www.elsevier.com/locate/conctc

# Estimation of treatment effects in weighted log-rank tests

Ray S. Lin*, Larry F. León

*Genentech Inc., South San Francisco, CA 94080, USA*

A B S T R A C T

Non-proportional hazards have been observed in clinical trials. The log-rank test loses power and the standard Cox model generally produces biased estimates under such conditions. Weighted log-rank tests have been utilized to increase the test power; however, it is not intuitive how to interpret the test result in terms of the clinical effect. We propose a Cox-model based time-varying treatment effect estimate to complement the weighted log-rank test.

The score test from the proposed model is equivalent to the weighted log-rank test, and a time-profile of the treatment effect can be obtained by fitting a time-varying covariate Cox model. Simulation results show that the proposed model preserves type-I error and achieve higher power than log-rank tests under non-proportional hazards scenarios. Whereas the standard Cox model produces biased effect estimates, the proposed model produces unbiased estimates if the weight function is correctly specified. It also achieves a better model fit and an enhanced flexibility to accommodate non-proportional hazards compared to the standard Cox model.

The proposed approach makes the assumptions of the weighted log-rank test explicit and the validity of assumptions can be assessed based on prior knowledge or model goodness-of-fit. It also helps to translate the weighted log-rank test results into quantitative estimates of the treatment effect with intuitive interpretation. The proposed method can be routinely conducted to complement weighted log-rank tests, especially in the setting where non-proportional hazards are expected.

## 1. Introduction

The log-rank test has been the most commonly used method for analyzing survival endpoints and is the most powerful under proportional hazards. The weighted log-rank test is its generalized form, which allows different weight assignment to time points and therefore is able to emphasize certain potion of the survival curves [1–3].

The weighted log-rank test has been utilized in studies with non-proportional hazards in order to increase power. For example, when a substantial portion of patients discontinue study treatments prematurely, the estimated treatment effect can be diluted and the power can be reduced since those patients may no longer derive benefit. By allocating higher weights to earlier time points, the test will focus on the earlier time period where there was limited treatment discontinuation and reflect the treatment benefit more accurately [4,5]. Similarly, some treatments may have a delayed period before exhibiting its full effect. In this case, lower weights can be allocated to earlier time points and thus focus the testing on the later time period [6–9].

Schoenfeld has shown that the most powerful weighted log-rank test is to assign the weights proportionally to the magnitude of log hazard ratio [10]. Type-I error is preserved if weights are pre-specified. The choice of weights can be based on prior knowledge, such as the characteristics of the treatment (e.g., delayed treatment effect, or long-term effect even after treatment discontinuation), the anticipated study design and conduct (e.g., cross-over, patient compliance, rate of early treatment discontinuation), and the general clinical context (e.g., long survival after treatment discontinuation, availability of non-protocol therapies or subsequent therapies).

Even though weighted log-rank tests have been in use, it is not intuitive how to examine the appropriateness of the weight function from the clinical perspective and how to interpret the test results in terms of treatment benefit.

We proposed a Cox-model based time-varying treatment effect estimate to complement the weighted log-rank test. This approach makes the assumptions of the weight function explicit in the form of relative magnitudes of treatment benefit over time, which can be examined and verified in the relevant clinical context. The score test of the proposed model is equivalent to the weighted log-rank test, and the estimate derived from the model provides a time-profile of the treatment effect. Prior to analyzing the data, the assumptions of the model (i.e., weight function or the relative magnitude of treatment effect over time) can be reviewed and examined based on prior knowledge. After model fitting,

* Corresponding author.
*E-mail address:* raylin@alumni.stanford.edu (R.S. Lin).

http://dx.doi.org/10.1016/j.conctc.2017.09.004
Received 23 May 2017; Received in revised form 9 September 2017; Accepted 13 September 2017
Available online 19 September 2017

the assumptions can be assessed based on model fit (e.g., through evaluating the patterns of residuals [11]). The estimated treatment effect (as a time-profile) can help clinical interpretation of the treatment benefit and facilitate the assessment on whether the benefit is clinically meaningful and economically valuable.

## 2. Method

### 2.1. Weighted log-rank test and cox model

Suppose $n$ patients are randomized into the treatment arm or the control arm in a clinical trial with a time-to-event endpoint. Let this treatment assignment be denoted by $X_1, X_2, ...X_n$, $X_i = 1$ if the i-th patient is assigned to the treatment arm and $X_i = 0$ otherwise. Let $T_1, T_2, ..., T_n$ denote the event or censoring times and $\delta_1, \delta_2, ..., \delta_n$ denote the status ($\delta_i = 1$ for an event and $\delta_i = 0$ if censored). Let $T_{(1)} \le T_{(2)} \le ...\le T_{(J)}$ denote the $J$ ordered event times. The weighted log-rank test statistics is

$$Z = \frac{\sum_{j=1}^{J} w(T_{(j)})(O_j - E_j)}{\sqrt{\sum_{j=1}^{J} w(T_{(j)})^2 V_j}}$$

where $O_j$ and $E_j$ denote the observed and expected (under null hypothesis) number of events in the treatment arm at time $T_{(j)}$, $V_j$ is the variance of $E_j$, and $w$ (.) is a weight function of time with non-negative value. Note that $Z$ stays the same if the weight function is multiplied or normalized by a constant scaler $k$.

For a Cox proportional hazards model, let the hazard function of the i-th patient be

$$\lambda(t; X_i) = \lambda_0(t) e^{w(t)\beta X_i}$$

where $\lambda_0(t)$ is the baseline hazard function and $\beta$ is the coefficient of the treatment. It can be shown that the score test of this Cox model is equivalent to the weighted log-rank test above (Appendix). Note that the weighted estimates proposed by Lin [12] and Sasieni [13] incorporate weights in the score function rather than in the hazard function. The score statistics derived from this proposed model is identical to their model, but the effect estimate $\hat{\beta}$ is different due to the different forms of the score function. Our approach provides an alternative method for estimating and interpreting the treatment effect under the weighted log-rank framework.

### 2.2. Effect adjustment factor

Given a weight function $w(t)$ in a weighted log-rank test, the effect adjustment factor $A(t)$ in the proposed model is defined as

$$A(t) = \frac{w(t)}{\max(w(t))}$$

so that $A(t)$ is non-negative and has the maximal value 1 at some time point(s).

The hazard function in the Cox model above can then be expressed as

$$\lambda(t; X) = \lambda_0 e^{A(t)\beta X}$$

Note that scaling by $\max(w(t))$ does not change the weighted log-rank statistics $Z$, and the score test of this model is still equivalent to the weighted log-rank test with weight function $w(t)$. The hazard function can also be viewed as a constant coefficient with a time-varying covariate $X^*(t) = A(t)X$, which represents the treatment assignment weighted by the adjustment factor. The coefficient $\beta$ can be easily estimated in the following Cox model once $X^*(t)$ is derived.

$$\lambda(t; X) = \lambda_0 e^{\beta X^*(t)}$$

The $\hat{\beta}$ estimated from models with time-varying covariates have been shown to be unbiased [14]. Because $A(t)$ is smaller than or equal to 1, $\beta$ represents the maximal effect in the time course, and the time points where the patients experience this maximal effect (i.e., at time $t$ with $A(t) = 1$) are assigned with the highest weights (i.e., at time $t$ with $w(t) = \max(w(t))$) in the corresponding weighted log-rank test. If the model is correct, this weighted log-rank test (and equivalently, the score test from this model) is optimal and will have the highest power based on Schoenfeld's proof [10].

The hazard ratio between the two arms $\mathrm{HR}(t)$ can then be derived as a function of time

$$HR(t) = \frac{\lambda_0 e^{A(t)\beta \times 1}}{\lambda_0 e^{A(t)\beta \times 0}} = e^{\beta A(t)} = [HR^F]^{A(t)}$$

where $\mathrm{HR}^F = e^{\beta}$ represents the full effect (i.e., maximal effect). The model incorporates the time-varying effect as the treatment coefficient $\beta$ weighted by the effect adjustment factor $A(t)$. Note that even though $\mathrm{HR}(t)$ seems to be a time-varying coefficient, the shape of this function is determined by $A(t)$ and only the magnitude $\beta$ is to be estimated from this Cox model. The score statistics (or equivalently, the weighted log-rank statistics) is testing whether the full effect $\mathrm{HR}^F$ is zero or not.

### 2.3. Examples of weight function and the corresponding effect adjustment factor

Various weight functions have been proposed for weighted log-rank tests. For example, the weight at time point $t$ can be assigned based on survival at $t$ (the Prentice-Wilcoxon or Peto-Peto test [1]), based on the number of patients at risk at $t$ (the Gehan-Breslow test [2,15]), or based on the proportions of patients who have discontinued study treatment at $t$ [4]. The $G^{\rho,\gamma}$ family proposed by Fleming and Harrington [3] is able to represent a variety of function shapes based on observed survival

$$w(t) = S(t)^{\rho}(1 - S(t))^{\gamma}$$

where $S(t)$ is the survival function of the pooled population; $\rho$ and $\gamma$ are parameters determining the shape of the weight function. The weighted log-rank test becomes the standard log-rank test when $\rho = \gamma = 0$ and becomes the Prentice-Wilcoxon test when $\rho = 1$ and $\gamma = 0$. The test allocates more weight at later time points when $\rho = 0$ and $\gamma = 1$ and more weight at the middle time points than the two ends when $\rho = 1$ and $\gamma = 1$.

Examples of these weight functions $w(t)$, the corresponding effect adjustment factors $A(t)$, and the time-varying hazard ratio $\mathrm{HR}(t)$ are presented in Fig. 1. An arbitrary survival curve $S(t)$ is generated for reference (represented by the dotted gray curves on the weight function panels).

In the standard log-rank test where the weight function is constant, the treatment is assumed to have the same level of effect (i.e., the full effect $\mathrm{HR}^F$) throughout the time course (i.e., $A(t) = 1$) whereas in the Prentice-Wilcoxon weight function, the treatment is assumed to have its full effect initially and then decreased monotonically in proportion to the observed survival rate (i.e., $\mathrm{HR}(t) = [\mathrm{HR}^F]^{S(t)}$). In the Fleming-Harrington $G^{1,1}$ weight function, treatment effect is assumed to have no effect prior to the first event and increase over time to reach its full effect around median survival time and then decrease over time (i.e., $\mathrm{HR}(t) = [\mathrm{HR}^F]^{S(t)(1-S(t))}$. In the weight function proposed by Lagakos and Bowden [4,5], treatment has the full effect initially and starts to decrease once patients start to discontinue study treatments (the “Decreasing Tail” scenario). In a hypothetical scenario where the treatment has delayed benefit, one can assume the treatment has minimal effect initially and then reaches its full effect after a certain delay period, illustrated in the right-most column.

### 2.4. Simulation studies

Three models and testing approaches were evaluated: (1) the standard Cox model with log-rank test, (2) our proposed model with weighted log-rank test, and (3) the short-term and long-term effect

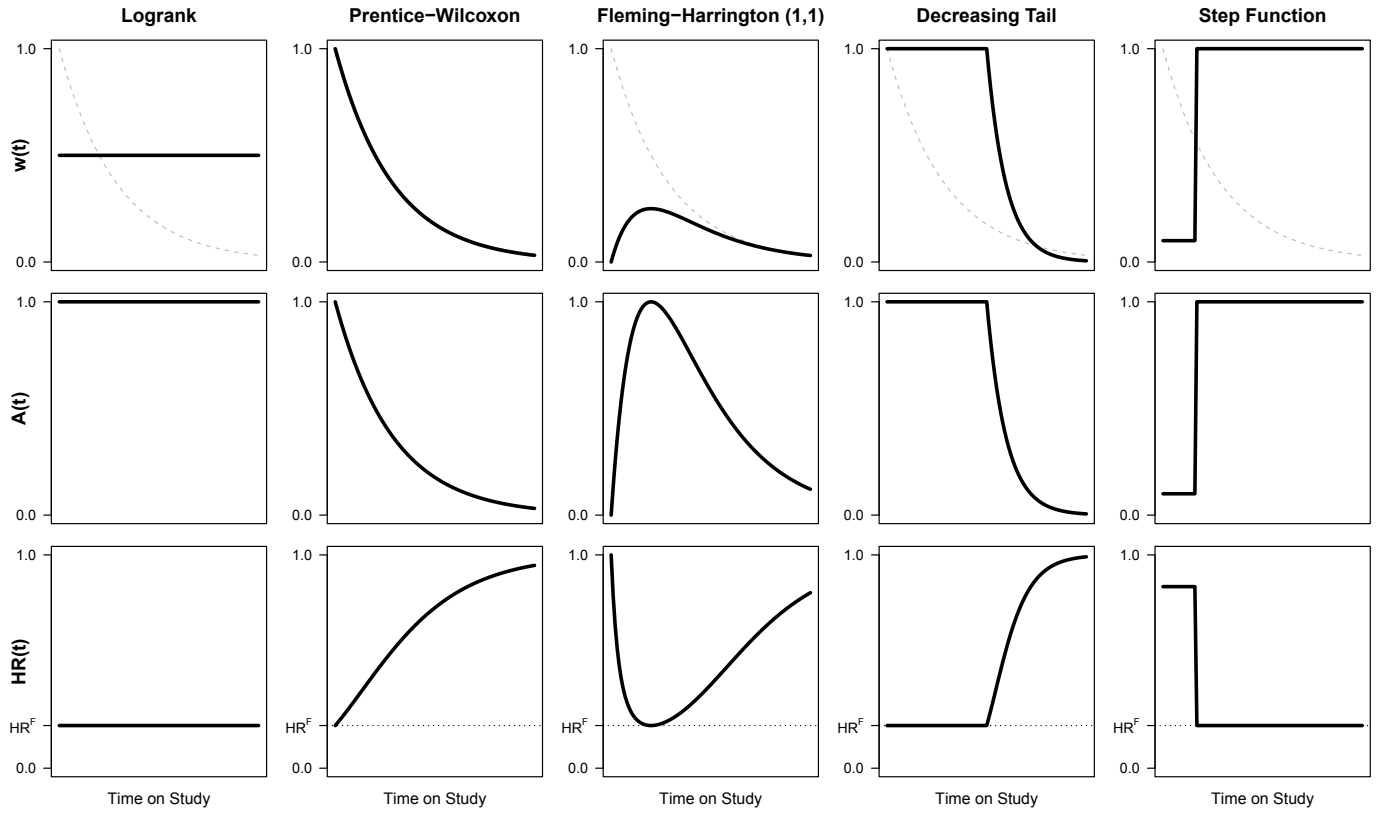


**Fig. 1.** Examples of weight functions, the corresponding effect adjustment factor $A(t)$, and the hazard ratio $\mathrm{HR}(t)$. The survival curve is represented by the dotted gray curves on the weight function panels.

model with adaptive weighted log-rank test by Yang and Prentice [16,17]. Two main settings were considered: (1) treatment has delayed effect and (2) the long-term treatment effect is reduced due to substantial treatment discontinuation. In each setting, several scenarios were simulated with 10,000 runs for each scenario to characterize the power, hazard ratio estimate, standard error, coverage of 95% confidence interval, and type-I error.

In the first setting, the treatment is assumed to have minimal effect (HR = 0.9) initially and will exhibit its full effect ($\mathrm{HR}^F = 0.68$) after a specific period of delay $\tau_D$. The study enrolls 400 patients and the analysis is triggered when 280 events have been observed, which provides about 90% power to detect the treatment's full effect (i.e., HR of 0.68) at a one-sided alpha level of 2.5%. The survival of the control arm follows an exponential distribution with a median of 6 months, and the survival of the treatment arm follows piece-wise exponential distribution with the piece-wise constant hazard determined by the hazard ratios above. The hazard for drop-out is assumed to be the same in the two arms and follow exponential distribution with an annual drop-out rate around 5%. The duration of the effect delay $\tau_D$ varies from 0 (no delay) to 6 months.

In the base case scenario (where the alternative hypothesis is assumed to be true), the effect adjustment factor is assumed to correctly reflect the true effect and be proportional to the log of the effect size: $A(t) = \log(0.9)/\log(0.68) = 0.27$ for $t < \tau_D$ and $A(t) = 1$ for $t \geq \tau_D$. This assignment corresponds to the weighted log-rank test that will give the highest power [10]. The same $A(t)$ function is used for type-I error evaluation under the null hypothesis. Additional sensitivity scenarios were also simulated to evaluate the effect if $A(t)$ is misspecified, if $\tau_D$ is misspecified, or if both $\tau_D$ and $A(t)$ are misspecified.

In the second setting, the treatment has its full effect $\mathrm{HR}^F$ initially and gradually decreases when more patients have discontinued from the treatment. The effect will eventually diminish after all the patients have discontinued their treatments. Let the proportion of patients who are still on treatment by time $t$ be $\gamma(t)$, then proportion of patients who discontinued treatment will be $1 - \gamma(t)$. If patients lose treatment effect completely immediately after treatment discontinuation, the hazard ratio at time $t$ will be $\mathrm{HR}(t) = \mathrm{HR}^F\gamma(t) + (1 - \gamma(t))$ [4]. However, the treatment may have a prolonged effect: that is, patients may still benefit from the treatment for a certain duration after their treatment discontinuation. Let the duration of this prolonged effect be $\tau_P$, then $\mathrm{HR}(t) = \mathrm{HR}^F\gamma(t - \tau_P) + (1 - \gamma(t - \tau_P))$, for any $t \geq \tau_P$; and $\mathrm{HR}(t) = \mathrm{HR}^F$ for any time $t < \tau_P$.

In the simulation, the study enrolls 720 patients with 510 events, which provides about 90% power to detect the treatment's full effect ($\mathrm{HR}^F = 0.75$) at a one-sided alpha level of 2.5%, assuming proportional hazards. The distribution for survival in the control arm and the drop-out in both arms are the same as the first setting, except that the median survival in the control is 12 months. The treatment discontinuation is assumed to follow exponential distribution with a 6-month median, and the prolonged effect $\tau_P$ is assumed to range from 0 to 24 months.

The effect adjustment factor $A(t)$ is assumed to be 1 for $t < \tau_P$ and $log\left[0.75 \times \gamma(t - \tau_P) + (1 - \gamma(t - \tau_P))\right]/log(0.75)$ for $t \geq \tau_P$ in the base case (alternative hypothesis) and in the null scenario. Sensitivity analyses evaluate model performance when $\tau_P$ is misspecified.

## 3. Results

### 3.1. Setting 1: delayed treatment effect

In the setting with delayed treatment effect, the power of the standard log-rank test (i.e., the score test of the standard Cox model), the weighted log-rank tests (i.e., the score test of the proposed model), and the Yang-Prentice tests all decrease as $\tau_D$ increases. However, the weighted log-rank test has consistently higher power than the other two tests when the weight function is correctly specified; the Yang-Prentice model is slightly more powerful than log-rank test but is not as powerful

**Table 1**
Setting 1 (Delayed treatment effect): characteristics of the log-rank test and the standard Cox model vs. the weighted log-rank test and the proposed model vs. the Yang-Prentice model (10,000 simulations per scenario).

| $\tau_D$ | Log-Rank and Standard Cox Model | | | | | Weighted Log-Rank and Proposed Model | | | | | Yang-Prentice Model[a] | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Power | $\hat{HR}$ | $se(\hat{\beta})$[b] | 95% CI coverage | Non-PH[c] | Power | $\hat{HR}$ | $se(\hat{\beta})$[b] | 95% CI coverage | Non-PH[c] | Power | $\hat{HR}_1$ | $se(\hat{\beta}_1)$[b] | $\hat{HR}_2$ | $se(\hat{\beta}_2)$[b] |
| $H_A$[d] | | | | | | | | | | | | | | | |
| 0 | 0.90 | 0.68 | 0.12 | 0.95 | 0.05 | 0.90 | 0.68 | 0.12 | 0.95 | 0.05 | 0.91 | 0.71 | 0.49 | 1.06 | 1.33 |
| 1 | 0.82 | 0.71 | 0.12 | 0.94 | 0.08 | 0.85 | 0.68 | 0.13 | 0.95 | 0.02 | 0.83 | 0.82 | 0.47 | 0.85 | 1.18 |
| 2 | 0.73 | 0.73 | 0.12 | 0.90 | 0.13 | 0.79 | 0.68 | 0.14 | 0.95 | 0.01 | 0.74 | 0.91 | 0.48 | 0.78 | 1.13 |
| 3 | 0.62 | 0.76 | 0.12 | 0.84 | 0.16 | 0.73 | 0.68 | 0.15 | 0.95 | 0.01 | 0.66 | 0.96 | 0.46 | 0.75 | 1.04 |
| 4 | 0.54 | 0.78 | 0.12 | 0.78 | 0.16 | 0.65 | 0.68 | 0.17 | 0.95 | 0.00 | 0.57 | 0.99 | 0.46 | 0.75 | 0.99 |
| 5 | 0.45 | 0.80 | 0.12 | 0.71 | 0.16 | 0.58 | 0.68 | 0.18 | 0.95 | 0.00 | 0.48 | 0.99 | 0.43 | 0.76 | 0.90 |
| 6 | 0.37 | 0.82 | 0.12 | 0.64 | 0.14 | 0.49 | 0.68 | 0.20 | 0.95 | 0.01 | 0.40 | 1.00 | 0.45 | 0.80 | 0.88 |
| $H_0$ | | | | | | | | | | | | | | | |
| 0 | 0.05 | 1.00 | 0.12 | 0.95 | 0.05 | 0.05 | 1.00 | 0.12 | 0.95 | 0.05 | 0.06 | 0.98 | 0.26 | 1.23 | 0.74 |
| 1 | 0.05 | 1.00 | 0.12 | 0.95 | 0.05 | 0.05 | 1.00 | 0.13 | 0.95 | 0.02 | 0.06 | 0.98 | 0.29 | 1.21 | 0.73 |
| 2 | 0.05 | 1.00 | 0.12 | 0.95 | 0.05 | 0.05 | 1.00 | 0.14 | 0.95 | 0.01 | 0.06 | 0.97 | 0.26 | 1.23 | 0.74 |
| 3 | 0.05 | 1.00 | 0.12 | 0.95 | 0.05 | 0.05 | 1.00 | 0.15 | 0.95 | 0.00 | 0.06 | 0.97 | 0.27 | 1.22 | 0.72 |
| 4 | 0.05 | 1.00 | 0.12 | 0.95 | 0.05 | 0.05 | 1.00 | 0.17 | 0.95 | 0.00 | 0.06 | 0.97 | 0.27 | 1.23 | 0.74 |
| 5 | 0.05 | 1.00 | 0.12 | 0.95 | 0.05 | 0.05 | 1.00 | 0.18 | 0.95 | 0.00 | 0.05 | 0.98 | 0.25 | 1.22 | 0.72 |
| 6 | 0.05 | 1.00 | 0.12 | 0.95 | 0.05 | 0.05 | 1.00 | 0.20 | 0.95 | 0.01 | 0.06 | 0.98 | 0.26 | 1.23 | 0.76 |

[a] $HR_1$ and $HR_2$ represent the short-term and the long-term effects in the Yang-Prentice model.
[b] Empirical standard error of $\log(HR)$
[c] Proportion of rejecting proportional hazards assumption at 5%.
[d] Alternative hypothesis: treatment hazard ratio is 0.9 during the delay period $\tau_D$ and is 0.68 thereafter.

as the weight log-rank test when $\tau_D$ increases (Table 1.).

In terms of hazard ratio estimates, the standard Cox model underestimates the effect and as $\tau_D$ increases, the bias increases and the conference interval coverage becomes incorrect. In contrast, the proposed model provides unbiased estimate and correct coverage across all scenarios. The standard error, however, is higher in the proposed model, especially when $\tau_D$ is longer (i.e., when a low weight is assigned for a longer time period). The empirical standard error is identical to the model's estimate for both models (not shown in the table).

The Yang-Prentice model provides estimates for the short-term and the long-term effects respectively. The treatment effect in this setting is captured by the short-term effect when $\tau_D$ is small (e.g., 0 month) and by the long-term effect when $\tau_D$ is large (e.g., 3 months or above). The estimates have bias in general, especially when $\tau_D$ is around 1–2 months. Since the two effects are estimated jointly, it may not be easy for the model to separate out the short-term versus the long-term effect when $\tau_D$ is not 0 but very small. The empirical standard errors of both estimates are substantially higher compared to the other two models.

Under the null hypothesis (HR = 1), both the standard and the proposed approaches preserve the type-I error and provide unbiased estimate of the treatment effect. The Yang-Prentice model in general preserves the type-I error yet the two estimates are slightly biased toward opposite directions. This could also be due to the joint estimation of the two effects and the model may pick up random patterns in the simulated data sets.

Note that when there is no delay, the weight function is a constant and thus $A(t) = 1$. The proposed approach becomes the standard log-rank test and the standard Cox model, and the two approaches have identical results.

Testing for non-proportionality [11] shows that the proportional hazards assumption is more likely to be violated in the standard Cox model (at 5% of alpha), especially when the delay is longer. In contrast, the proposed Cox model is able to mitigate the violation by incorporating the time-varying covariate based on $A(t)$.

Sensitivity analysis (Fig. 2) shows that weighted log-rank test has lower power (the left panel) when $\tau_D$ is misspecified (true $\tau_D$ is 3 months). However, its power is still higher than the standard Cox model (i.e., 63%). In the proposed model, the treatment effect (the middle panel) is underestimated when the model assumes a shorter delay than the actual $\tau_D$ (3 months). This is similar to the standard Cox model (which assumes no delay) because it tries to average the effect over the delay period where the effect is lower, which biases the overall effect estimate toward null. On the other hand, the effect is overestimated when the assumed delay is longer than 3 months: the model tries to average out the effect over the assumed delay period (which actually includes some period with full effect), resulting in bias away from null.

The hazard ratio plot (the right panel) shows the time-profile of the treatment effect: the true profile is represented by the blue curve ($\tau_D$ is 3 months); the red curve represents the standard Cox model (assuming $\tau_D$ is 0 and constant hazard ratio); the purple curve represent the model assuming $\tau_D$ is 6 months. The overestimated HR 0.65 in the purple curve lasted for a longer period (6 months) before the treatment exhibits its full effect; in other words, the patients have limited benefit for a long period based on this model (longer than actual). This time-profile reflects the model assumptions explicitly. Therefore, even though the full effect size is overestimated, a relatively balanced assessment of the overall benefit across the whole time course can be examined by the time-profile plot. Whether this benefit profile is clinically meaningful and economically justifiable can then be assessed under this clear and explicit framework.

When $\tau_D$ is correctly specified (i.e., 3 months) yet the adjustment factor $A(t)$ is misspecified during the first 3 months (Fig. 3), the power and point estimate appear to be similar to the correct model (i.e., $A(t) = 0.27$) when $A(t)$ is not much different from the correct model. The higher $A(t)$ is (i.e., assuming strong effect during the delay period), the worse the model performance is: the extreme case with $A(t) = 1$ is equivalent to the standard Cox model and has the lowest power and greatest bias.

Fig. 4 shows the results when both $\tau_D$ and $A(t)$ are misspecified: the true $\tau_D$ is 1.5 months, and the $A(t)$ is 0.27 during the first 1.5 months and is 1 afterwards; yet the model assumes $\tau_D = 3$, with $A(t)$ during the delay period ranging from 0 to 1. Because $\tau_D$ is short, hazards are proportional most of the time and hence Cox model is able to achieve 77% power. The proposed model in general achieve comparable or higher power than the Cox model; however, when it puts too little weight during the delay period (e.g., $A(t) \leq 0.2$), its power could be even slightly lower than Cox model. This demonstrates again that power decreases due to the higher standard error when a low weight is assigned for a long period. The hazard ratio estimate from the Cox model still has higher bias than the proposed model, even when the

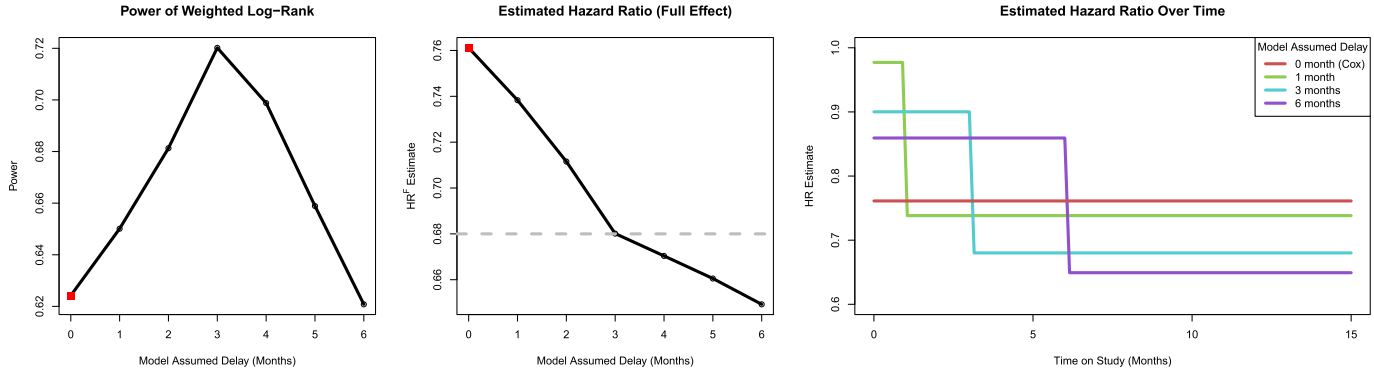


**Fig. 2.** Characteristics of the weighted log-rank test and proposed model when the prolonged effect $\tau_D$ is misspecified. The true $\tau_D$ is 3 months, and the true hazard ratio is 0.68. † Power and hazard ratio estimate from the Cox model are shown by the red squares on the left and the middle panels.

model is misspecified. Furthermore, the estimate from the proposed model is relatively stable across various $A(t)$ values, and thus the model misspecification seems to have limited impact on the hazard ratio estimate.

### 3.2. Setting 2: reduced long-term treatment effect

In the second setting with a reduced treatment effect in the long-term (Table 2), the powers of both the log-rank and weighted log-rank tests decrease when $\tau_P$ is shorter. Similar to Setting 1, the weighted log-rank test has consistently higher power than the log-rank test.

The hazard ratio estimated from Cox model is biased toward null and it is more likely to violate the proportional hazards assumption, especially with shorter $\tau_P$, because the proportional hazards assumption implicitly assumes $\tau_P$ is infinity (i.e., the effect is constant and prolonged throughout the study regardless of treatment discontinuation). In contrast, the proposed model provides an unbiased estimate and correct coverage across all scenarios and have lower incidence of violating proportional hazards assumption. Standard error estimated by the model (not shown in the table) are the same as the empirical data, and the proposed approach has higher standard error when the model assumes a shorter $\tau_P$ (i.e., more time points are assigned with lower weights).

When $\tau_P$ is very long (e.g., 24 months), treatment effect is almost constant (proportional hazards) throughout the study. Therefore, treatment discontinuation has minimal impact on power, and the two approaches show almost identical results because most of the time points are assigned with a weight of 1. As in Setting 1, both approaches preserve the type-I error.

In the Yang-Prentice model, the short-term effect captures the treatment effect in Setting 2. Its power is between the proposed model and the Cox model, similar trend as in Setting 1. The hazard ratio estimate fluctuates across scenarios around the truth (i.e., 0.75) with high standard error. Under the null hypothesis, the model preserves type-I error and has slightly biased estimates.

When the weight function is misspecified (Fig. 5, true $\tau_P$ is 8 months), the power is reduced but is relatively stable and is higher than the standard Cox model (66%). The treatment effect is underestimated when the model assumed a longer $\tau_P$ than the truth (8 months) and overestimated when the assumed a shorter $\tau_P$. The hazard ratio time-profile shows the true profile ($\tau_P$ = 8 months) in light blue; the standard Cox model in red. When the model assumes $\tau_P$ = 0, the hazard ratio is biased (0.68); however the overestimated effect (hazard ratio <0.75) lasts for only a short period of time (during the first 3 months) and effect size is in fact biased toward null for all the time points after 3 months.

## 4. Discussion

The log-rank test has been widely used in survival analysis and is generally the gold-standard approach. It implicitly assumes a constant treatment effect over time and is the most powerful when such a condition is met. Consequently, the corresponding Cox model (i.e., the model with the score test equivalent to the log-rank test) explicitly assumes constant treatment effect (or proportional hazards) and therefore the effect estimate derived from the model is constant over time. However, when the proportional hazards assumption is violated, the log-rank test has reduced power even though it is still a valid test in this setting; on the other hand, the Cox model is no longer (strictly) valid and thus the effect estimated from the model is biased (depending on the nature of non-proportionality).

Non-proportional hazards have been observed in clinical trials, and the simulations conducted in this research were designed to reflect some of such cases: one example of the delayed treatment effect is in

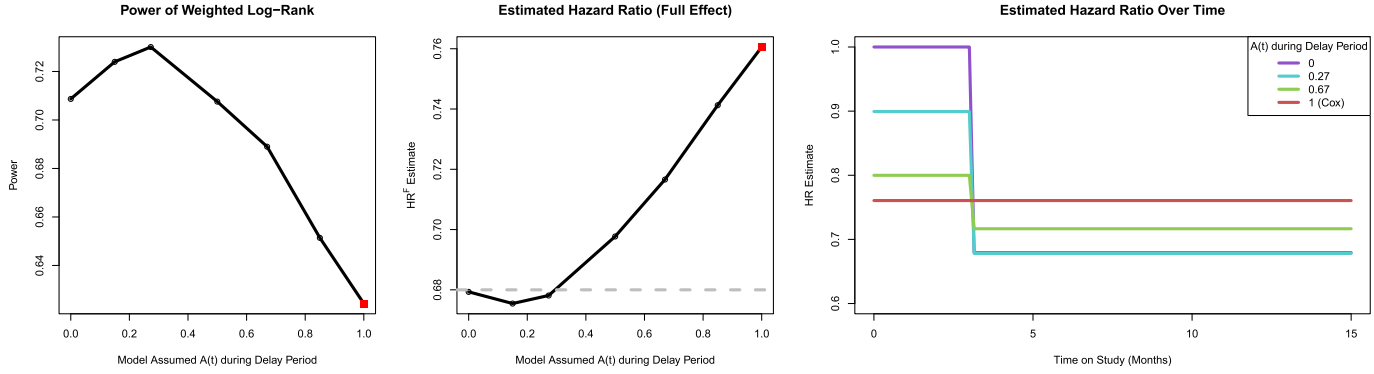


**Fig. 3.** Characteristics of the weighted log-rank test and proposed model when the adjustment factor $A(t)$ during the delay period (i.e., the first 3 months) is misspecified. The true hazard ratio is 0.9 during the first 3 months and 0.68 afterwards. The correct $A(t)$ is thus 0.27 and 1 respectively. † Power and hazard ratio estimate from the Cox model are shown by the red squares on the left and the middle panels. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

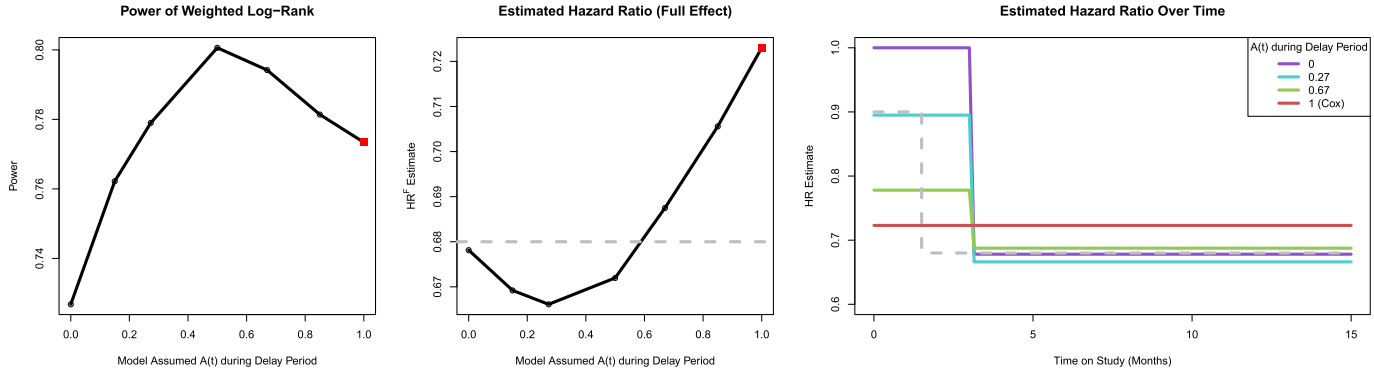


**Fig. 4.** Characteristics of the weighted log-rank test and proposed model when both $\tau_D$ and $A(t)$ are misspecified. The true $\tau_D$ is 1.5 months, and the correct $A(t)$ is 0.27 during the first 1.5 months and is 1 afterwards (represented in gray dashed line). The model assumes $\tau_D$ is 3 months, with $A(t)$ ranging from 0 to 1. † Power and hazard ratio estimate from the Cox model are shown by the red squares on the left and the middle panels.

vaccine or immunotherapy, where a certain period is needed for the immune system to react and respond to the treatment; therefore limited effect is observed initially and the full effect will be observed once the immune system is fully activated [7,9]. The scenario of the reduced long-term effect is also commonly observed in oncology studies with overall survival as the primary endpoint. Patients in the studies usually discontinue the assigned treatment once they experienced disease progression. They may then receive any non-protocol treatments post disease progression, which will confound the overall survival. For example, patients assigned to the control arm may receive the study treatment or similar treatments as standard care, through cross-over, off-label use, or participation of other clinical trials. This may still result in diluted treatment effect even if patients discontinued from the treatment arm are actually experiencing prolonged treatment benefit.

The Weighted log-rank test is a generalized method of the log-rank. Correspondingly, the proposed approach is a generalized method of the standard Cox model. The Weighted log-rank test and the proposed Cox model relax the assumption of a constant effect (or proportional hazards): it allows the treatment effect to vary over time and assumes relative magnitudes of the effect as a function of time via the adjustment factor $A(t)$. By the pre-specified weight function $w(t)$ in the weighted log-rank test, the proposed model assumes a specific shape of the treatment effect over time and the magnitude of the full effect is then estimated based on such assumption. Allowing time-varying effects can address non-proportional hazards in Cox model and provide a more unbiased effect estimate.

From the perspective of weighted log-rank test, the standard log-rank is a special case where equal weights are allocated to all time points. Its corresponding Cox model assumes the effect is a horizontal line over time (i.e., constant effect) and then estimates the effect magnitude based on this straight-line assumption. Intuitively, the Cox model is estimating the treatment effect averaged over all time points, and the corresponding log-rank test is testing whether such averaged effect is statistically significant or not.

When there are non-proportional hazards, the weight function can be introduced to the weighted log-rank to mitigate power loss, and similarly, the corresponding time-varying effect can be incorporated into Cox model to improve model fit and provide less biased estimate [14]. The effect estimate from the proposed Cox model is essentially derived from: (1) the average effect magnitude estimated from the time points with the highest weight (i.e., the time points with $A(t) = 1$) and (2) the adjusted effect magnitdude estimated from each of other time point j

**Table 2**
Setting 2 (Reduced long-term treatment effect): characteristics of the log-rank test and the standard Cox model vs. the weighted log-rank test and the proposed model vs. the Yang-Prentice model.

| $\tau_P$ | Log-Rank and Standard Cox Model | | | | | Weighted Log-Rank and Proposed Model | | | | | Yang-Prentice Model[a] | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Power | $\hat{HR}$ | $se(\hat{\beta})$[b] | 95% CI coverage | Non-PH[c] | Power | $\hat{HR}$ | $se(\hat{\beta})$[b] | 95% CI coverage | Non-PH[c] | Power | $\hat{HR}_1$ | $se(\hat{\beta}_1)$[b] | $\hat{HR}_2$ | $se(\hat{\beta}_2)$[b] |
| $H_A$[d] | | | | | | | | | | | | | | | |
| 0 | 0.29 | 0.88 | 0.09 | 0.55 | 0.14 | 0.38 | 0.75 | 0.17 | 0.95 | 0.00 | 0.33 | 0.77 | 0.21 | 1.28 | 0.58 |
| 2 | 0.38 | 0.86 | 0.09 | 0.64 | 0.17 | 0.50 | 0.75 | 0.15 | 0.95 | 0.00 | 0.44 | 0.74 | 0.22 | 1.34 | 0.67 |
| 4 | 0.49 | 0.84 | 0.09 | 0.74 | 0.18 | 0.60 | 0.75 | 0.13 | 0.95 | 0.00 | 0.53 | 0.72 | 0.29 | 1.39 | 0.78 |
| 6 | 0.58 | 0.83 | 0.09 | 0.81 | 0.18 | 0.68 | 0.75 | 0.12 | 0.95 | 0.00 | 0.63 | 0.72 | 0.30 | 1.42 | 0.92 |
| 8 | 0.66 | 0.81 | 0.09 | 0.85 | 0.15 | 0.73 | 0.75 | 0.11 | 0.95 | 0.01 | 0.69 | 0.72 | 0.34 | 1.44 | 1.02 |
| 10 | 0.71 | 0.80 | 0.09 | 0.89 | 0.13 | 0.77 | 0.75 | 0.11 | 0.95 | 0.01 | 0.75 | 0.74 | 0.41 | 1.44 | 1.14 |
| 12 | 0.77 | 0.79 | 0.09 | 0.91 | 0.11 | 0.81 | 0.75 | 0.10 | 0.95 | 0.01 | 0.79 | 0.74 | 0.40 | 1.38 | 1.16 |
| 24 | 0.89 | 0.75 | 0.09 | 0.95 | 0.05 | 0.89 | 0.75 | 0.09 | 0.94 | 0.05 | 0.91 | 0.78 | 0.43 | 1.01 | 1.06 |
| $H_0$ | | | | | | | | | | | | | | | |
| 0 | 0.05 | 1.00 | 0.09 | 0.95 | 0.05 | 0.05 | 1.00 | 0.17 | 0.95 | 0.00 | 0.05 | 0.99 | 0.19 | 1.08 | 0.37 |
| 2 | 0.05 | 1.00 | 0.09 | 0.95 | 0.05 | 0.05 | 1.00 | 0.14 | 0.95 | 0.00 | 0.05 | 0.99 | 0.18 | 1.07 | 0.35 |
| 4 | 0.05 | 1.00 | 0.09 | 0.95 | 0.05 | 0.05 | 1.00 | 0.12 | 0.95 | 0.00 | 0.05 | 0.98 | 0.18 | 1.07 | 0.34 |
| 6 | 0.05 | 1.00 | 0.09 | 0.95 | 0.05 | 0.05 | 1.00 | 0.11 | 0.95 | 0.00 | 0.05 | 0.99 | 0.19 | 1.07 | 0.34 |
| 8 | 0.05 | 1.00 | 0.09 | 0.95 | 0.05 | 0.05 | 1.00 | 0.11 | 0.95 | 0.01 | 0.05 | 0.98 | 0.18 | 1.07 | 0.35 |
| 10 | 0.05 | 1.00 | 0.09 | 0.95 | 0.05 | 0.05 | 1.00 | 0.10 | 0.95 | 0.02 | 0.05 | 0.98 | 0.18 | 1.07 | 0.34 |
| 12 | 0.05 | 1.00 | 0.09 | 0.95 | 0.05 | 0.05 | 1.00 | 0.10 | 0.95 | 0.02 | 0.06 | 0.98 | 0.18 | 1.07 | 0.35 |
| 24 | 0.05 | 1.00 | 0.09 | 0.95 | 0.05 | 0.05 | 1.00 | 0.09 | 0.95 | 0.05 | 0.05 | 0.98 | 0.18 | 1.07 | 0.35 |

[a] $HR_1$ and $HR_2$ represent the short-term and the long-term effects in the Yang-Prentice model.
[b] Empirical standard error of $\log(HR)$
[c] Proportion of rejecting proportional hazards assumption at 5%.
[d] Alternative hypothesis: treatment hazard ratio is 0.75 initially and the effect reduced in proportion to treatment discontinuation proportion, after a duration of prolonged effect $\tau_P$.

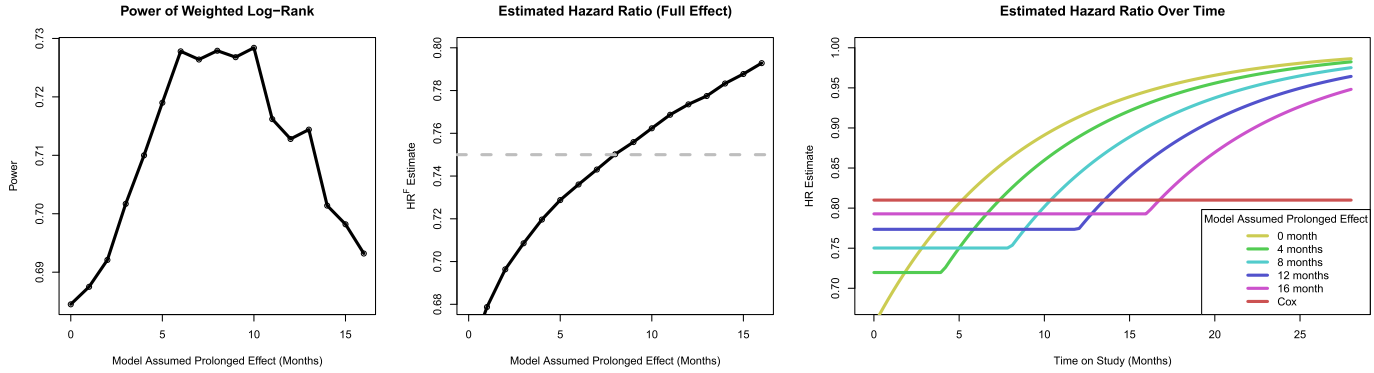


**Fig. 5.** Characteristics of the weighted log-rank test and proposed model when the prolonged effect $\tau_P$ is misspecified. The true $\tau_P$ is 8 months, and the true hazard ratio is 0.75.

(where $A(j) < 1$) based on discounting the true effect by A(j). This adjustment allows the model to reflect non-proportional hazards and to capture and estimate the time-varying treatment effect.

The standard error of the effect estimate is in general higher than the standard Cox model, consistent with the findings from Lin's and Sasieni's approach [12,13]; however, the proposed model still achieves relatively higher power because its estimate is unbiased if the weight function is correctly specified. Even if the weight function is misspecified, the proposed model will lose power yet still has higher power and produces less biased estimate than the standard Cox model in the simulated scenarios.

One possible way to reduce the standard error is to assign high weights ($A(t)$ close to 1) for a long period. A low weight ($A(t)$ close 0) implies that all the observations at the corresponding time points are almost treated as censored [13]. This pseudo censoring increases the standard error and reduces power substantially, especially when the low weights are assigned for a substantially long time period. Assigning the highest weight for a longer period will not only reduce the pseudo censoring but will also average out the full effect across more time points (where $A(t) = 1$) and improve the precision of the effect estimate. However, the trade-off is that it may not reflect the nuance of the non-proportional hazards at certain time periods and will introduce bias. The standard Cox model is the extreme case where it assigns the same weight to all time points and achieves the lowest standard error yet suffers from potential bias. The choice of weight function is essentially a trade-off between bias and variance, and this could be guided by prior knowledge.

A wide range of other methods have been proposed to test survival curves and to estimate treatment effects in the presence of non-proportional hazards. These methods estimate additional model parameters from the data and model the time-varying effect via various model structures and estimation approaches, such as the Yang-Prentice model used in our simulations, which has a short-term and a long-term effect [16,17], or using a cubic spline [18], penalized partial likelihood [19], kernel-weighted partial likelihood [20,21], adaptive group lasso [22], and via piece-wise constant hazard ratio models with change-point detection [23–25]. These models enable flexible structures that can be tailored to the data.

These models are summarized as a spectrum in Table 3. The standard Cox model and log-rank test assume the shape of the hazard ratio time-profile (i.e., a straight line over time) and then estimate and test the magnitude of the effect. The proposed model and the weighted log-rank generalize the assumption by allowing the effect time-profile to take different shapes through the specification of $w(t)$ or $A(t)$ and then estimate and test the full effect $HR^F$. Other flexible methods take an even more generalized approach by estimating both the magnitude and the shape of the effect based on data. Whereas these flexible approaches are able to describe the data more accurately and achieve better model fit, they might be fitting to random data patterns that occurred only in this dataset and are not consistent with prior knowledge nor can be reproduced in future studies. This is consistent with our finding that the Yang-Prentice model tends to have substantially higher standard error for the hazard ratio estimates and may sometimes have slight bias. Type I error may also be inflated due to the data-driven nature of these models [26]. On the other end of the spectrum, the log-rank test and standard Cox model are robust to such random noise and preserve type-I error, but they are subject to bias if the proportional hazards assumption is not met.

**Table 3**
Spectrum of models for effect estimation.

| Model | Magnitude of the effect | Shape of the effect (relative magnitude over time) |
|---|---|---|
| Standard Cox model | data | pre-specified as a straight line (constant effect) |
| Proposed model | data | pre-specified (time-varying effect) |
| Flexible models | data | data (time-varying effect) |

The proposed model is a balance between the two ends: it is not as data-driven as the flexible models, yet it allows incorporation of prior knowledge to mitigate non-proportional hazards via pre-specified adjustment factor $A(t)$. Because $A(t)$ is pre-specified, the type-I error is preserved, as demonstrated in the simulations. The proposed approach makes $w(t)$ in weighted log-rank test explicit so that both $w(t)$ and $A(t)$ can be reviewed and examined prior to model fitting based on whether it is biologically reasonable and whether it can reflect the characteristics of the treatment, the study design and conduct, and the general clinical context including treatment landscape and clinical practice. The simulations also suggest that the hazard ratio estimate and power are relatively robust if the model (change-point) is misspecified.

After model fitting, the assumption can be verified by model fitness (e.g., through assessing the patterns of residuals [11]). The effect time-profile HR($t$) reflects the assumption explicitly and can help to assess whether the effect is substantial and meaningful from the clinical perspective and is valuable and justifiable from the health economics' perspective. For example, in the scenario with delayed treatment benefit, it may not be effective to treat patients who are not likely to survive beyond the delay period. Similarly, in the scenario with reduced long-term effect, the payers may not plan to reimburse the treatment beyond the time point when the treatment effect is expected to diminish.

This is an alternative approach to the log-rank and the standard Cox model that can potentially describe the treatment effect in a more accurate manner. As suggested by Sasieni [13], weighted log-rank can be routinely performed in addition to the standard methods. However, the weighted log-rank test only examines whether the full effect $HR^F$ is significant, and it is also important to know when the treatment starts to reach its full effect, how long it lasts, and when and how much it decreases or changes over time. The proposed time-profile approach helps to translate the weighted log-rank tests to quantitative effect estimates,

which facilitates the assessment of treatment effect in terms of its clinical meaningfulness and economical values. The proposed method can be routinely conducted to complement weighted log-rank test, especially in the setting where non-proportional hazards are expected.

**Acknowledgements**

The authors thank the three anonymous reviewers for their valuable comments, which have led to substantial improvement on this paper.

## A. Appendix

For a clinical trial with $n$ patients randomized into a treatment arm or a control arm, let the treatment assignment be denoted by $X_1, X_2, \ldots X_n$, $X_i = 1$ if the i-th patient is assigned to the treatment arm and $X_i = 0$ otherwise. Let $T_1, T_2, \ldots, T_n$ denote the event or censoring times and $\delta_1, \delta_2, \ldots \delta_n$ denote the status ($\delta_i = 1$ for an event and $\delta_i = 0$ if censored). Let $T_{(1)} \le T_{(2)} \le \ldots \le T_{(J)}$ denote the $J$ ordered event times and $X_{(1)}, X_{(2)}, \ldots, X_{(J)}$ denote the corresponding treatment assignment.

The weighted log-rank test statistics is

$$Z = \frac{\sum_{j=1}^{J} w(T_{(j)})(O_j - E_j)}{\sqrt{\sum_{j=1}^{J} w(T_{(j)})^2 V_j}}$$

where $O_j$ and $E_j$ denote the observed and expected (under null hypothesis) number of events in the treatment arm at time $T_{(j)}$, $V_j$ is the variance of $E_j$, and $w(.)$ is weight function of time with non-negative value.

We propose a Cox proportional hazards model with the hazard function of the i-th patient:

$$\lambda(t; X_i) = \lambda_0(t) e^{w(t)\beta X_i}$$

where $\lambda_0(t)$ is the baseline hazard function and $\beta$ is the coefficient of the treatment.

The partial likelihood of the model $L(\beta)$ is

$$L(\beta) = \prod_{j=1}^{J} \frac{\lambda_0(t) e^{w\left(T_{(j)}\right)\beta X_{(j)}}}{\sum_{l \in R_j} \lambda_0(t) e^{w\left(T_{(j)}\right)\beta X_l}} = \prod_{j=1}^{J} \frac{e^{w\left(T_{(j)}\right)\beta X_{(j)}}}{\sum_{l \in R_j} e^{w\left(T_{(j)}\right)\beta X_l}}$$

where $R_j$ is the set of patients at risk at time $T_{(j)}$.

The log partial likelihood $l(\beta)$, the score function $U(\beta)$, the information matrix $I(\beta)$, and the score statistics $U(0)/\sqrt{I(0)}$ are as following:

$$l(\beta) = \sum_{j=1}^{J} \left\{ w(T_{(j)}) \beta X_{(j)} - \log \sum_{l \in R_j} e^{w(T_{(j)})\beta X_l} \right\}$$

$$U(\beta) = \frac{\partial l(\beta)}{\partial \beta} = \sum_{j=1}^{J} \left\{ w(T_{(j)}) X_{(j)} - \frac{\sum_{l \in R_j} w(T_{(j)}) X_l e^{w\left(T_{(j)}\right)\beta X_l}}{\sum_{l \in R_j} e^{w\left(T_{(j)}\right)\beta X_l}} \right\}$$

$$I(\beta) = \frac{\partial^2 l(\beta)}{\partial \beta^2} = \sum_{j=1}^{J} \frac{\left(\sum_{l \in R_j} w(T_{(j)})^2 X_l^2 e^{w\left(T_{(j)}\right)\beta X_l}\right)\left(\sum_{l \in R_j} e^{w\left(T_{(j)}\right)\beta X_l}\right)}{\left(\sum_{l \in R_j} e^{w\left(T_{(j)}\right)\beta X_l}\right)^2} - \sum_{j=1}^{J} \frac{\left(\sum_{l \in R_j} w(T_{(j)}) X_l e^{w\left(T_{(j)}\right)\beta X_l}\right)\left(\sum_{l \in R_j} w(T_{(j)}) X_l e^{w\left(T_{(j)}\right)\beta X_l}\right)}{\left(\sum_{l \in R_j} e^{w\left(T_{(j)}\right)\beta X_l}\right)^2}$$

$$= \sum_{j=1}^{J} \left\{ \frac{\sum_{l \in R_j} w(T_{(j)})^2 X_l^2 e^{w\left(T_{(j)}\right)\beta X_l}}{\sum_{l \in R_j} e^{w\left(T_{(j)}\right)\beta X_l}} - \left\{ \frac{\sum_{l \in R_j} w(T_{(j)}) X_l e^{w\left(T_{(j)}\right)\beta X_l}}{\sum_{l \in R_j} e^{w\left(T_{(j)}\right)\beta X_l}} \right\}^2 \right\}$$

$$\frac{U(0)}{\sqrt{I(0)}} = \sum_{j=1}^{J} \left\{ w(T_{(j)}) X_{(j)} - \frac{\sum_{l \in R_j} w(T_{(j)}) X_l}{\sum_{l \in R_j} 1} \right\} \Bigg/ \sqrt{\sum_{j=1}^{J} \left\{ \frac{\sum_{l \in R_j} w(T_{(j)})^2 X_l^2}{\sum_{l \in R_j} 1} - \left\{ \frac{\sum_{l \in R_j} w(T_{(j)}) X_l}{\sum_{l \in R_j} 1} \right\}^2 \right\}}$$

$$= \sum_{j=1}^{J} \left\{ w(T_{(j)}) X_{(j)} - \frac{\sum_{l \in R_j} w(T_{(j)}) X_l}{|R_j|} \right\} \Bigg/ \sqrt{\sum_{j=1}^{J} \left\{ \frac{\sum_{l \in R_j} w(T_{(j)})^2 X_l^2}{|R_j|} - \left\{ \frac{\sum_{l \in R_j} w(T_{(j)}) X_l}{|R_j|} \right\}^2 \right\}}$$

$$= \frac{\left(\sum_{j=1}^{J} [w(T_{(j)}) O_{(j)} - w(T_{(j)}) E_{(j)}]\right)}{\left(\sqrt{\sum_{j=1}^{J} [w(T_{(j)})^2 E_{(j)} - w(T_{(j)})^2 E_{(j)}^2]}\right)} = \frac{\left(\sum_{j=1}^{J} w(T_{(j)}) [O_{(j)} - E_{(j)}]\right)}{\left(\sqrt{\sum_{j=1}^{J} w(T_{(j)})^2 E_{(j)} (1 - E_{(j)})}\right)}$$

$$= \frac{\left(\sum_{j=1}^{J} w(T_{(j)}) [O_{(j)} - E_{(j)}]\right)}{\left(\sqrt{\sum_{j=1}^{J} w(T_{(j)})^2 V_{(j)}}\right)}$$

which is the weighted log-rank statistic $Z$.

## References


[1] R.L. Prentice, Linear rank tests with right censored data, Biometrika 65 (1) (1978) 167–179.
[2] E.A. Gehan, A generalized Wilcoxon test for comparing arbitrarily singly-censored samples, Biometrika 52 (1965) 203–223.
[3] D.P. Harrington, T.R. Fleming, A class of rank test procedures for censored survival data, Biometrika 69 (3) (1982) 553–566.
[4] S.W. Lagakos, L.L. Lim, J.M. Robins, Adjusting for early treatment termination in comparative clinical trials, Stat. Med. 9 (Dec 1990) 1417–1424 discussion 1433–7.
[5] J. Bowden, S. Seaman, X. Huang, I.R. White, Gaining power and precision by using model-based weights in the analysis of late stage cancer trials with substantial treatment switching, Stat. Med. 35 (9) (2016) 1423–1440.
[6] D.M. Zucker, E. Lakatos, Weighted log rank type statistics for comparing survival curves when there is a time lag in the effectiveness of treatment, Biometrika 77 (4) (1990) 853–864.
[7] B. Zhu, Y. Park, B. Zhen, Z. Xu, Designing therapeutic cancer vaccine trials with delayed treatment effect, Stat. Med. 36 (4) (2017) 592–605.
[8] G.D. Fine, Consequences of delayed treatment effects on analysis of time-to-event endpoints, Drug Inf. J. 41 (2007).
[9] T.-T. Chen, Statistical issues and challenges in immuno-oncology, J. Immunother. Cancer 1 (1) (2013) 18.
[10] D. Schoenfeld, The asymptotic properties of nonparametric tests for comparing survival distributions, Biometrika 68 (Dec 1981) 219–316.
[11] P. Grambsch, T. Therneau, Proportional hazards tests and diagnostics based on weighted residuals, Biometrika 81 (3) (1994) 515–526.
[12] D.Y. Lin, Goodness-of-fit analysis for the Cox regression model based on a class of parameter estimators, J. Amer. Stat. Assoc. 86 (415) (1991) 725–728.
[13] P. Sasieni, Maximum weighted partial likelihood estimators for the Cox model, J. Amer. Stat. Assoc. 88 (1993) 144–152.
[14] P.K. Andersen, R.D. Gill, Cox's regression model for counting processes: a large sample study, Ann. Stat. (1982) 1100–1120.
[15] N. Breslow, A generalized kruskal-wallis test for comparing k samples subject to unequal patterns of censorship, Biometrika 57 (1) (1970) 579–594.
[16] S. Yang, R. Prentice, Semiparametric analysis of short-term and long-term hazard ratios with two-sample survival data, Biometrika 92 (2005) 1–17.
[17] S. Yang, R. Prentice, Improved logrank-type tests for survival data using adaptive weights, Biometrics 66 (2010) 30–38.
[18] K.R. Hess, Assessing time-by-covariate interactions in proportional hazards regression models using cubic spline functions, Stat. Med. 13 (May 1994) 1045–1062.
[19] D.M. Zucker, A.F. Karr, Nonparametric survival analysis with time-dependent covariate effects: a penalized partial likelihood approach, Ann. Stat. 18 (1) (1990) 329–353.
[20] Z. Cai, Y. Sun, Local linear estimation for time-dependent coefficients in Cox's regression models, Scand. J. Stat. 30 (1) (2003) 93–111.
[21] L. Tian, D. Zucker, L.J. Wei, On the Cox model with time-varying regression coefficients, J. Amer. Stat. Assoc. 100 (469) (2005) 172–183.
[22] J. Yan, J. Huang, Model selection for Cox models with time-varying coefficients, Biometrics 68 (2) (2012) 419–428.
[23] K.Y. Liang, S.G. Self, X.H. Liu, The cox proportional hazards model with change point: an epidemiologic application, Biometrics 46 (Sep 1990) 783–793.
[24] M. Liu, W. Lu, Y. Shao, A Monte Carlo approach for change-point detection in the Cox proportional hazards model, Stat. Med. 27 (19) (2008) 3894–3909.
[25] P. He, L. Fang, Z. Su, A sequential testing approach to detecting multiple change points in the proportional hazards model, Stat. Med. 32 (7) (2013) 1239–1245.
[26] C. Chauvel, J. O'Quigley, Tests for comparing estimated survival functions, Biometrika 101 (3) (2014) 535–552.