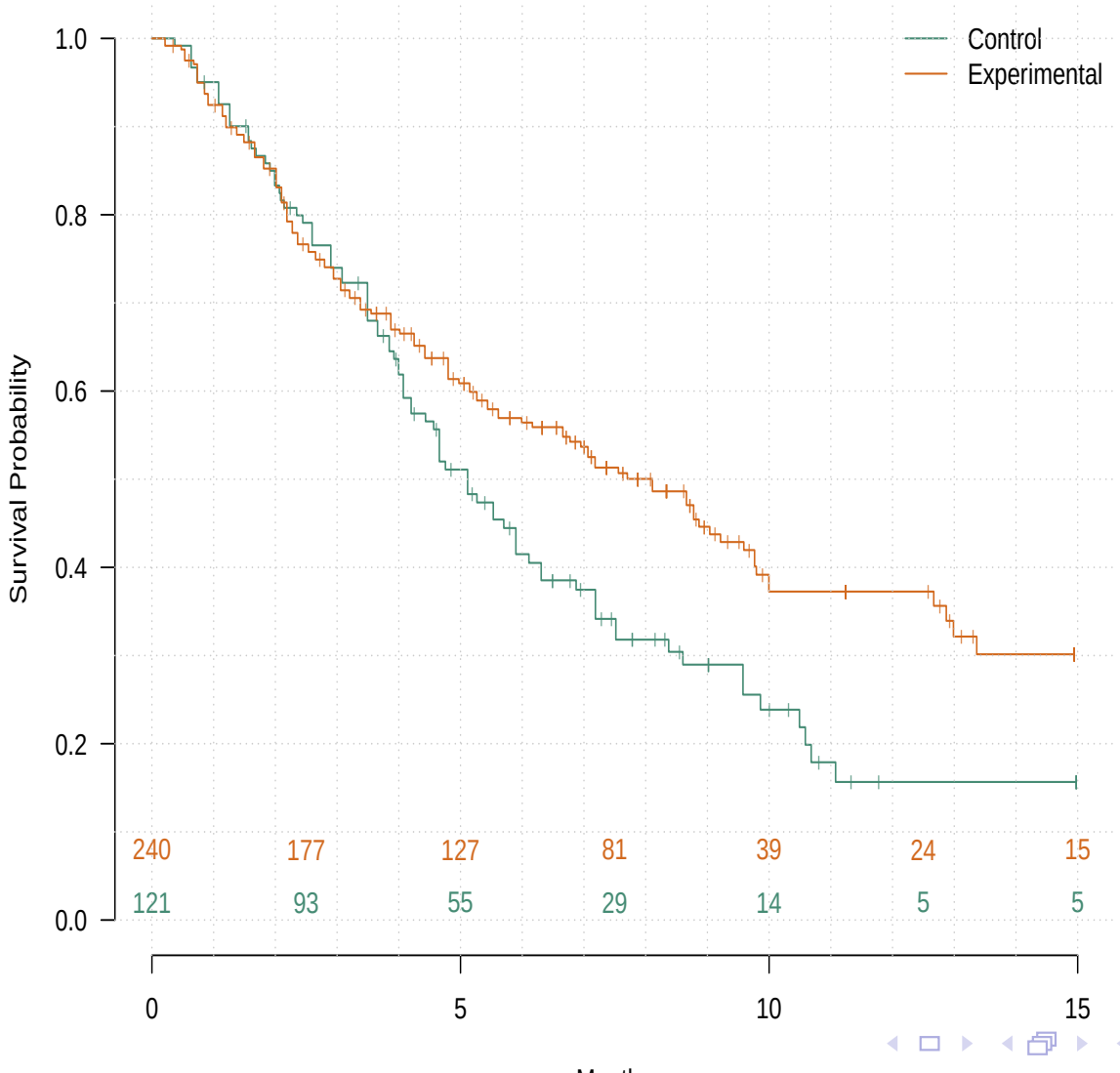
A weighted log-rank test and associated effect estimator for cancer trials with delayed treatment effect

Chang Yu

Department of Biostatistics
Vanderbilt University Medical Center

PSI Journal Club, March 24, 2021

Nivolumab on head and neck cancer, Overall Survival, Re-constructed data from Ferris et al. (2016)



Setup and notation

- Randomize n subjects into two treatment groups ($X_j = 0$: control arm and $X_j = 1$: experimental arm, $j = 1, ..., n$).
- D is the set of subjects who experienced the event.
- t_j is the event time or censoring time for the j^{th} subject and we assume the event times are distinct.
- Let $n_i(t)$ be the number of subjects at risk for the event before time t for treatment group i.
- $p(t) = n_1(t)/\{n_0(t) + n_1(t)\}$

Motivated by Schoenfeld (1981) *Biometrika*

- The test statistic

$$S = \frac{\sum_{j \in D} w_j \left(X_j - p(t_j)\right)}{\left[\sum_{j \in D} w_j^2 p(t_j)(1 - p(t_j))\right]^{1/2}} \tag{1}$$

- The standard log-rank test when $w_j = 1$.
- The Fleming-Harrington test (Fleming & Harrington, 1991) when

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

where $\rho \geq 0$, $\gamma \geq 0$ and $\widehat{S}(t)$ is the pooled estimate of the survival function at time t.

A hazard ratio model

- The hazard ratio (HR)

$$\lambda(t) = h_1(t)/h_0(t) = \begin{cases} 1 & t \le t_1 \\ \dfrac{\lambda - 1}{t_2 - t_1}(t - t_1) + 1 & t_1 < t \le t_2 \\ \lambda & t > t_2 \end{cases} \tag{2}$$

- $h_0(t)$ and $h_1(t)$ are the hazard functions of the control and the experimental groups respectively.
- Discussed by clinicians in cancer immunotherapy research (Hoos et al. 2010, JNCI, and others.)

Weight functions

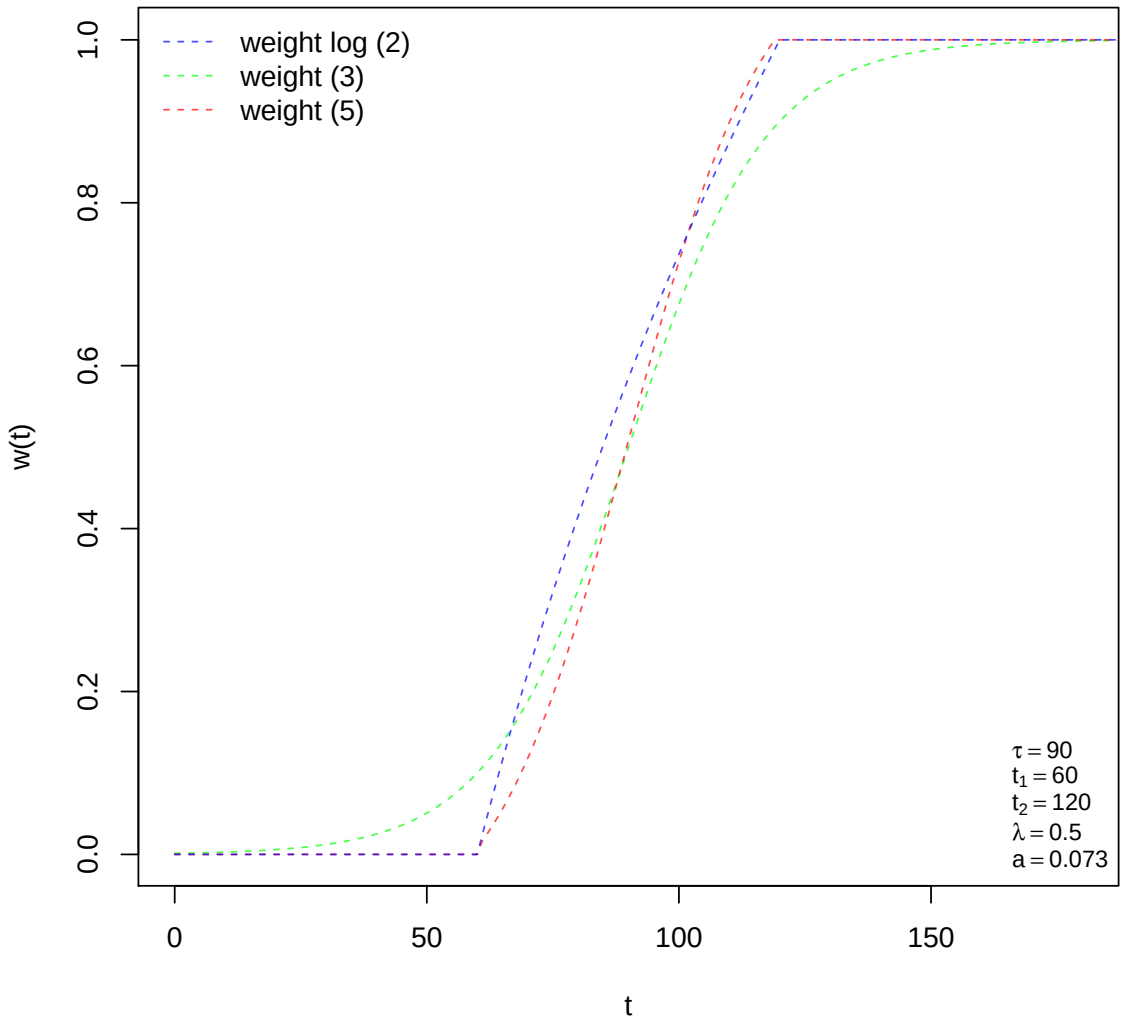
- Set weight w_1 to w_2 at time t_1 and t_2

$$w(t) = \frac{e^{a(t-\tau)}}{1 + e^{a(t-\tau)}} \tag{3}$$

- Motivated by Schoenfeld (1981) and Xu et al. (2017, Stat Med), the weighted log-rank test (1) with weight proportional to the logarithm of the HR at the event time would asymptotically maximize its power.
-

$$w_a(t) = \begin{cases} 0 & t \leq t_1 \\ \dfrac{w(t) - w(t_1)}{w(t_2) - w(t_1)} & t_1 < t \leq t_2 \\ 1 & t > t_2 \end{cases} \tag{4}$$

Three weight functions



Test statistic

Theorem Test statistic (1) with weight functions (3) and (4) is asymptotically normally distributed with mean μ and unit variance.

Schoenfeld approximation of μ (Schoenfeld, 1981, Biometrika) using the Taylor expansion when $\log(h_1(t)/h_0(t)) \sim O(n^{-1/2})$,

$$\mu = \frac{n^{1/2} \int w(t) \log(h_1(t)/h_0(t)) \pi(t)(1-\pi(t)) V(t) dt}{[\int (w(t))^2 \pi(t)(1-\pi(t)) V(t) dt]^{1/2}} \qquad (5)$$

Test statistic

- The integration is over the range from 0 to ∞;
- $$V(t) = P_0 f_0(t)(1 - H_0(t)) + P_1 f_1(t)(1 - H_1(t))\,;$$
- $$\pi(t) = \frac{P_1(1 - F_1(t))(1 - H_1(t))}{P_0(1 - F_0(t))(1 - H_0(t)) + P_1(1 - F_1(t))(1 - H_1(t))}\,.$$

Sample size and power

- The key is to assess (analytically or numerically)

$$\mu = \frac{n^{1/2} \int w(t) \log(h_1(t)/h_0(t)) \pi(t)(1-\pi(t)) V(t) dt}{[\int (w(t))^2 \pi(t)(1-\pi(t)) V(t) dt]^{1/2}} = n^{1/2} R$$

- R programs to numerically evaluate R.
- Sample size

$$n = [(Z_{1-\alpha/2} + Z_{1-\beta})/R]^2$$

- Power

$$1 - \beta = \Phi(\mu - Z_{1-\alpha/2}) + \Phi(-\mu - Z_{1-\alpha/2})$$

Estimation through a connection between weighted log-rank test and weighted Cox regression

If we use weight (3) or (4) in the weighted Cox regression (WCR)

- Our weighted log-rank test is the score test from the weighted Cox regression.
- $\exp(\hat{\beta})$ obtained from WCR with censoring correction, using weight $w(t)\widehat{G}(t)^{-1}$, provides an estimate of the average hazard ratio (AHR).
- Schemper et al. (2009, Stat Med) discussed how AHR could be estimated in connection with WCR.

Average hazard ratio (AHR)

- We compare three AHR's as estimands of the treatment effect in our study.
- The AHR-CR is estimated using uniform one weight with censoring correction.
- The AHR-WCR is estimated using the Prentice weight $S(t)$ with censoring correction.
- The WCR using weights (3) and (4) show a similar performance so we focus on the latter. The estimator is denoted as AHR-WCR2.

Simulation algorithm

1. n subjects are randomized with $1:1$ ratio to the two arms. Generate subjects' enrollment times U from a uniform distribution with rate $1/A$, A is the enrollment period.
2. For subjects in the control arm, their event time T_0 follows an exponential (h_0) distribution.
3. For subjects in the experimental arm, their event time T_1 could be
   - Under the null: type I error rate is controlled.
   - Under various delayed scenarios.

Simulation algorithm

4. Then we have the observed survival time $Z = min\{T, B - U\}$ and the event indicator $\delta = I\{T \leq B - U\}$, where $T = T_0 \cup T_1$. We assume the cause to loss-of-follow-up is administrative censoring.
5. Apply the proposed weighted log-rank tests using weights (3) and (4), the standard log-rank test, or tests in the Fleming-Harrington $G^{\rho,\gamma}$ class.
6. Repeat steps 1 through 5 for 10,000 simulation replicates to evaluate the empirical type I error rate or power.

Empirical power for 3 transition periods

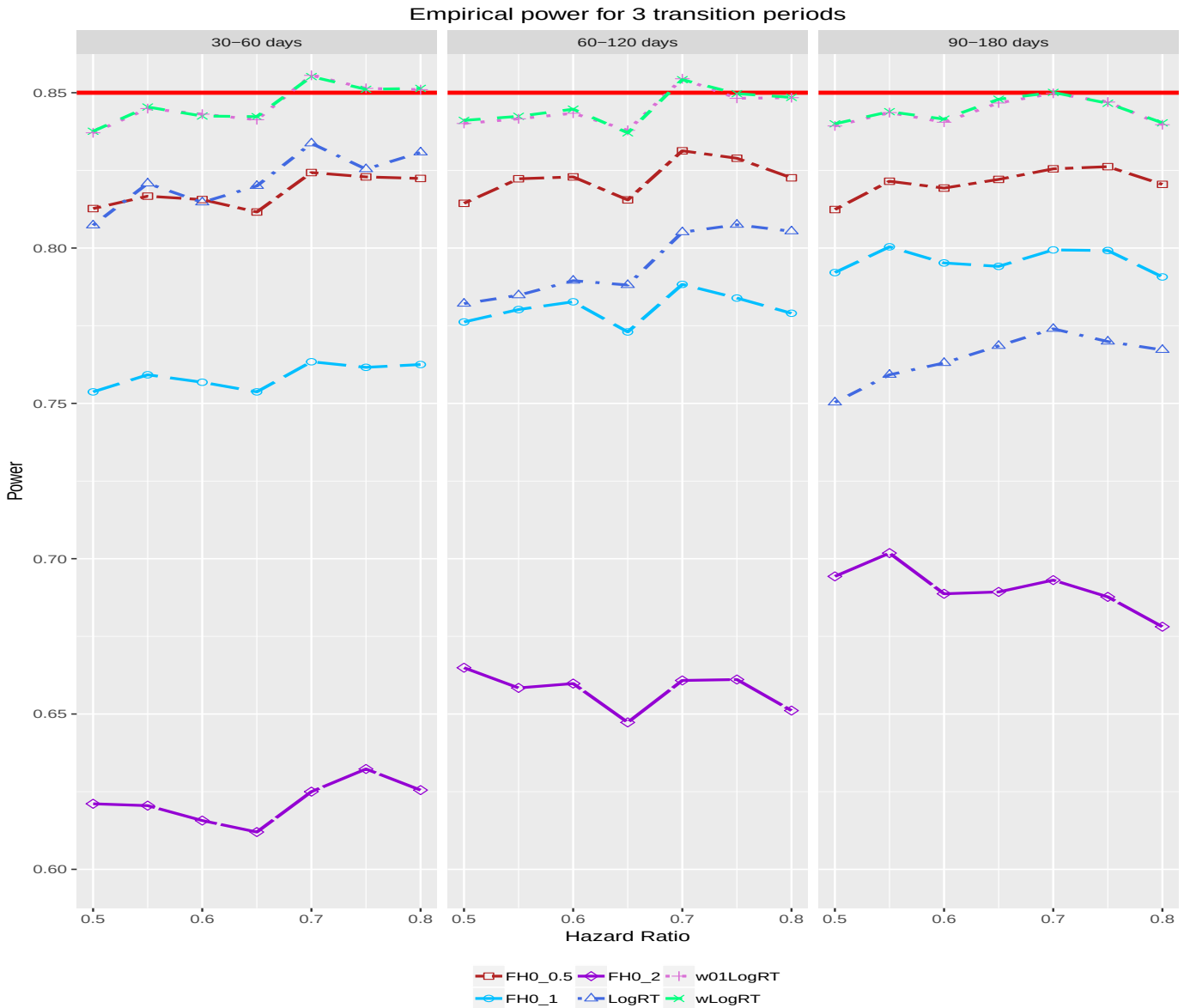


Table 1: Empirical power of 5 tests: wLogRT using weight (3), w01LogRT using weight (4), 3 tests in the $G^{\rho,\gamma}$ class with $(\rho = 0, \gamma = 0.5)$ (FH0_0.5), $(\rho = 0, \gamma = 1)$ (FH0_1), and $(\rho = 0, \gamma = 2)$ (FH0_2), and the standard log-rank test (LogRT).

Transition period used	Center (days)	Width (days)	Sample size	Empirical power(%)					
				wLogRT	w01LogRT	FH0_0.5	FH0_1	FH0_2	LogRT
Correct center, correct width									
90 - 180	135	90	232	85.0	84.8	82.4	80.7	71.4	74.9
Wrong center, correct width									
0 - 90	45	90	198	73.2	72.8	75.6	74.1	64.0	67.5
45 - 135	90	90	214	79.9	79.7	78.8	77.0	67.8	71.2
135 - 225	180	90	252	86.3	86.2	85.1	83.6	74.5	77.6
180 - 270	225	90	276	87.9	87.5	88.3	87.3	78.5	82.2
Correct center, wrong width									
125 - 145	135	20	226	83.6	83.4	80.8	79.1	70.0	73.9
110 - 160	135	50	228	84.3	84.3	81.8	80.2	70.7	74.6
40 - 230	135	190	240	85.4	85.6	83.4	81.7	73.2	75.9
Wrong center, wrong width									
30 - 70	50	40	196	72.5	72.1	74.9	72.9	63.7	67.4
0 - 100	50	100	200	74.3	73.8	76.0	74.3	65.1	68.1
200 - 240	220	40	268	86.5	86.2	87.4	85.9	77.8	80.6
170 - 270	220	100	274	87.7	87.4	87.9	86.5	78.1	81.2

Note: The sample size is calculated using the NESA method for the weighted log-rank tests to have 85% power to detect HR 0.5 under various specifications of the transition period. Simulation set-up: there is a delayed treatment effect with the transition period 90-180 days (centered at 135 days with width 90 days); the enrollment period is $A = 1$ year and the maximum follow-up is $B = 3$ years; the control group hazard rate is 0.31 (equivalently survival rate 40% at the end of year 3); nominal $\alpha = 0.05$ is used; the number of simulation replicates is 10,000.

Table 2: Re-analysis of the overall survival data from the trial of nivolumab, Ferris et al. (2016)

Transition period	P-values		AHR-WCR2	
	wLogRT	w01LogRT	Estimate	CI
0 - 4	0.000419	0.000330	0.605	(0.405-0.804)
1 - 3	0.000396	0.000458	0.602	(0.4-0.804)
2 - 2	0.002319	0.002646	0.639	(0.424-0.854)
0 - 5	0.000331	0.000225	0.593	(0.388-0.799)
1 - 4	0.000192	0.000145	0.576	(0.372-0.779)
1.5 - 3.5	0.000152	0.000137	0.569	(0.365-0.773)
2 - 3	0.000132	0.000133	0.561	(0.356-0.766)
2.5 - 2.5	0.000121	0.000124	0.555	(0.35-0.76)
2 - 4	0.000083	0.000059	0.542	(0.334-0.751)
3 - 3	0.000087	0.000087	0.538	(0.327-0.749)
2 - 5	0.000136	0.000104	0.547	(0.325-0.769)
3.5 - 3.5	0.000050	0.000056	0.531	(0.308-0.755)
LogRT		0.006976		
AHR-CR			0.685	(0.489-0.881)
AHR-WCR			0.731	(0.533-0.928)

Discussion

- The regularity condition $\log(h_1(t)/h_0(t)) \sim O(n^{-1/2})$ under which Schoenfeld (1981) derived the Schoenfeld approximation (10) does not appear to be stringent in practice.
- Usually the true transition period is not known in practice. Investigators should lean toward later-centered, wider transition period to be conservative when they design a trial.
- Further research on treatment effect estimator is needed.
- Software: we have R programs to implement our methods.

Acknowledgment

- The work is a collaboration with Xiang Huang and Hui Nian of Vanderbilt University Medical Center, and Philip He of AstraZeneca.

- This work was supported in part by Vanderbilt CTSA grant UL1 TR000445 from NIH/NCATS, R01 CA149633 from NIH/NCI, R21 HL129020, PPG HL108800, R01HL111259, R21HL123829 from NIH/NHLBI, R01HS022093 from NIH/AHRQ (CY), and R21 HL129020, R01CA202936 (XH), and R01CA202936 (HN).

Thank You!

References

- Schoenfeld D (1981). The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika* **68**:316-319.
- Fleming TR, Harrington DP (1991). Counting processes and survival analysis. John Wiley & Sons: New York, 1991.
- Xu Z, Zhen B, Park Y, Zhu B (2017). Designing therapeutic cancer vaccine trials with delayed treatment effect. *Statistics in Medicine* **36**(4):592-605.
- Hoos A, Eggermont AMM, et al. (2010). Improved endpoints for cancer immunotherapy Trials. *JNCI: Journal of the National Cancer Institute* **102**(18):1388-1397.
- Schemper M, Wakounig S, Heinze G (2009). The estimation of average hazard ratios by weighted Cox regression. *Statistics in Medicine* **28**:2473-2489.
- Breslow NE, Edler L and Berger J (1984). A two-sample censored-data rank test for acceleration. *Biometrics* **40**4:1049-1062.