Survival Analysis using a 5-STAR Approach in Randomized Clinical Trials

Devan V. Mehrotra* and Rachel Marceau West
Biostatistics and Research Decision Sciences

* devan_mehrotra@merck.com

ASA Webinar
May 26, 2020

Outline

- Introduction
- Motivating example
- Logrank test
- Alternatives to logrank test
- Proposed 5-STAR approach
- Other examples
- Simulation results
- Conclusions

Introduction

- Randomized clinical trial, two treatment arms (A=test, B=control)

  Random variable T_j = survival time under treatment *j*

- $S_j(t) = Pr(T_j > t)$

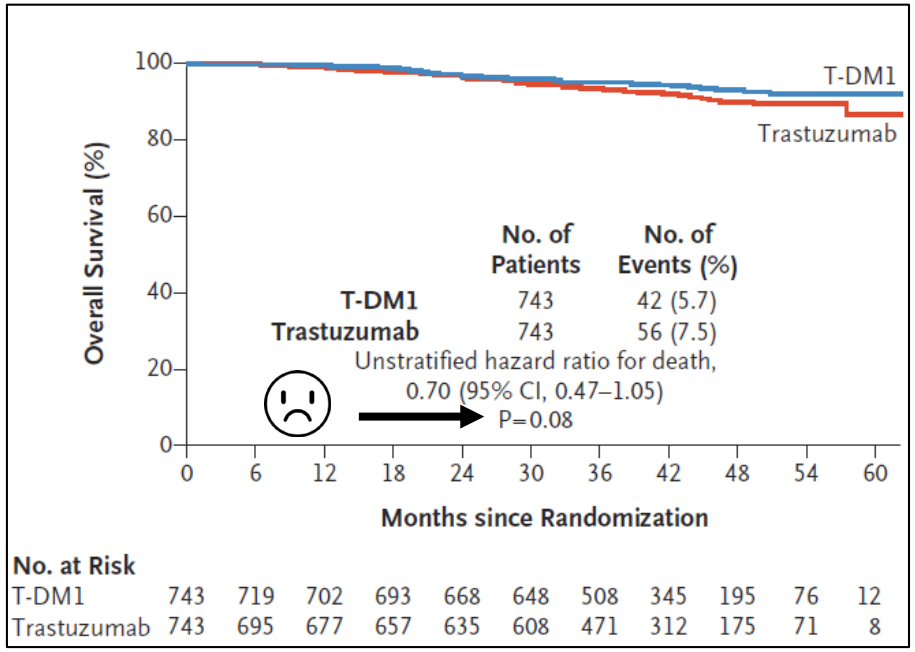
  $H_0: S_A(t) = S_B(t)$ for all t

- Statistical deliverables:
  1. P-value associated with test of H_0
  2. Point estimate and CI for an **interpretable** population level treatment effect parameter (estimand)

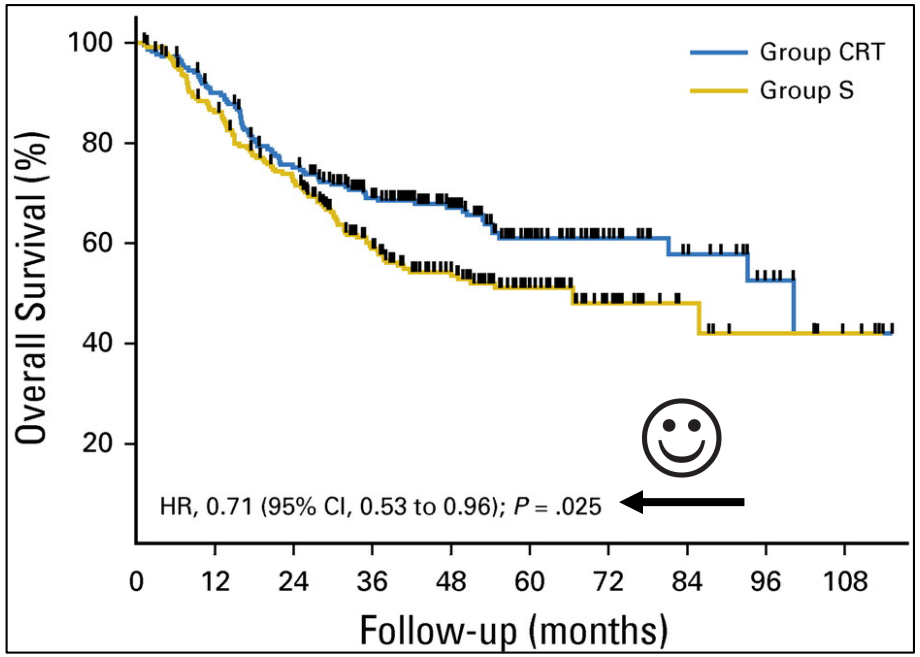
  Ideally, (1) and (2) should be aligned, per ICH E9/R1 (2019)

Logrank Test – Popular Analysis in RCTs

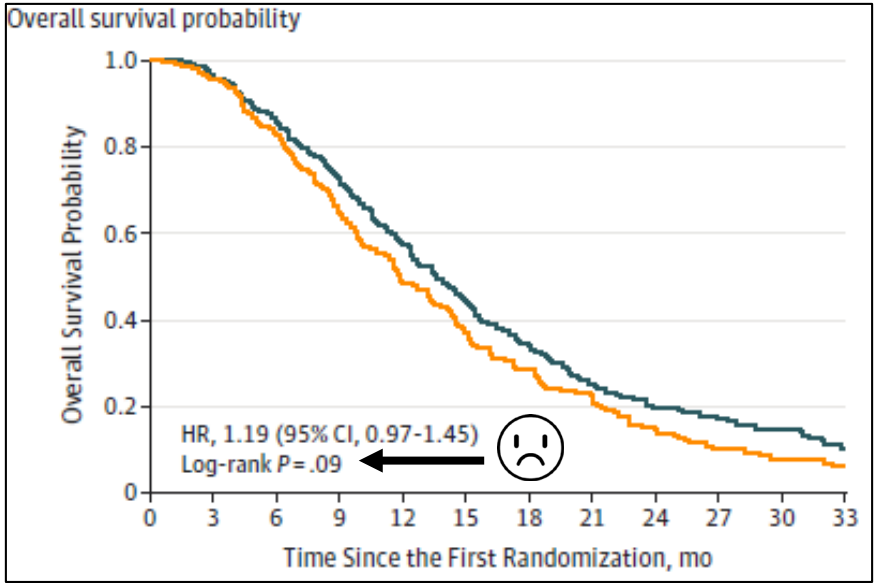
NEJM 2018



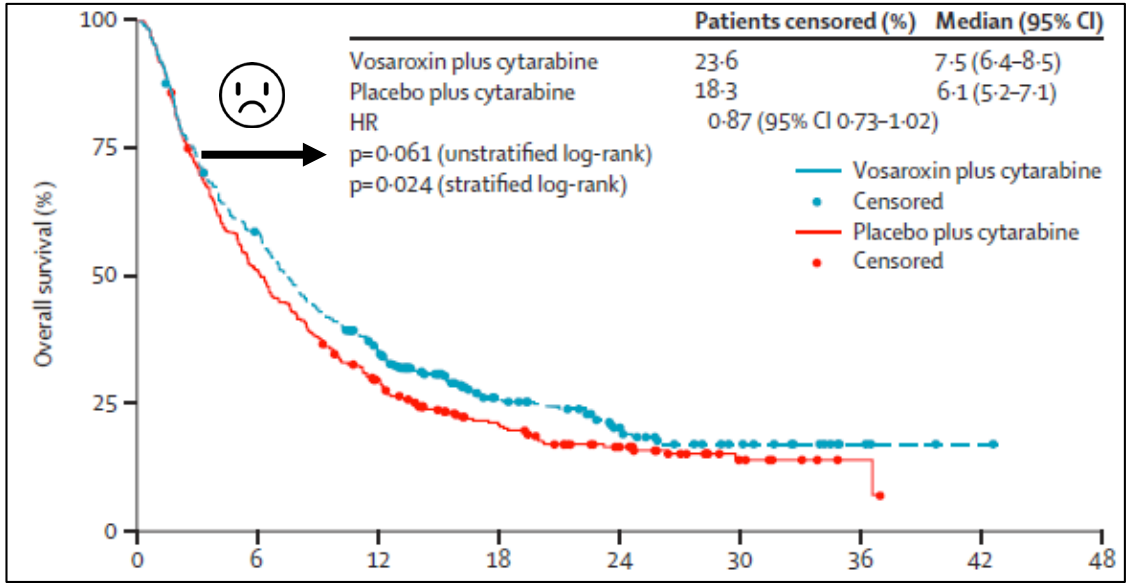
JCO 2018



JAMA 2016



Lancet Oncology 2015



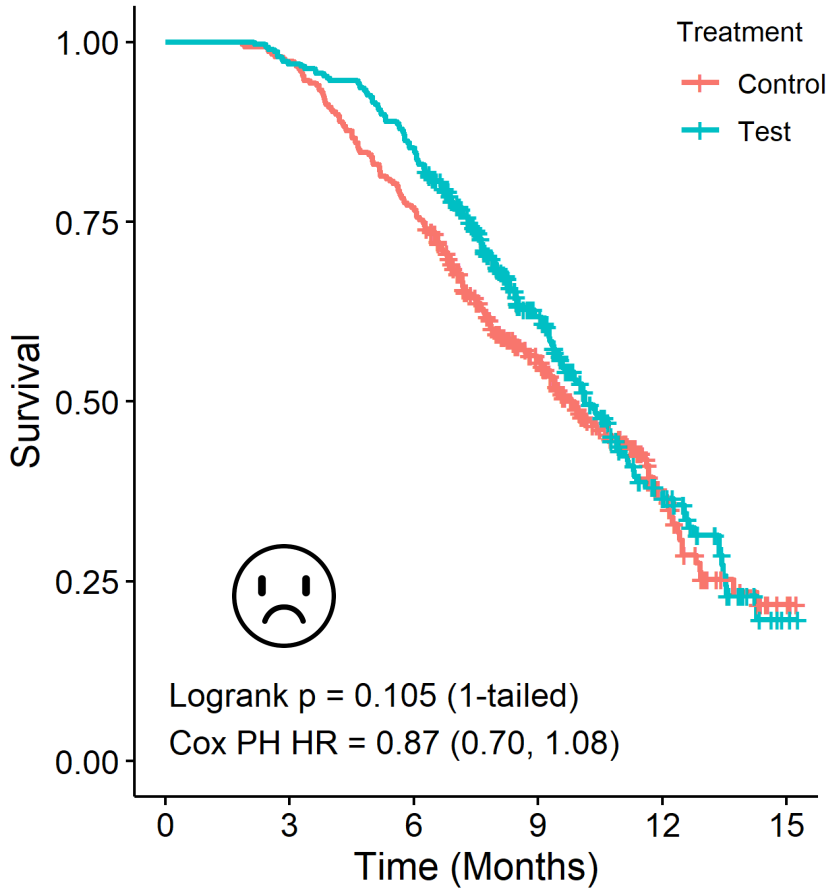
Logrank Test (continued)

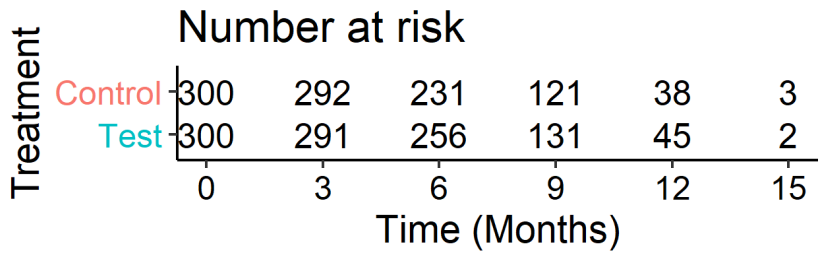
- Logrank test = score test from the Cox proportional hazards (PH) model
- When the hazard functions for A and B are **proportional**
  - Logrank test is optimal for testing H_{null}
  - $\theta(t) = \frac{\log\{S_A(t)\}}{\log\{S_B(t)\}} = \theta$ for all t
  - θ is the time-invariant hazard ratio (HR)
- When the hazard functions for A and B are **not proportional**
  - Logrank test is no longer optimal (potential power loss)
  - The Cox PH model HR estimate can be hard to interpret

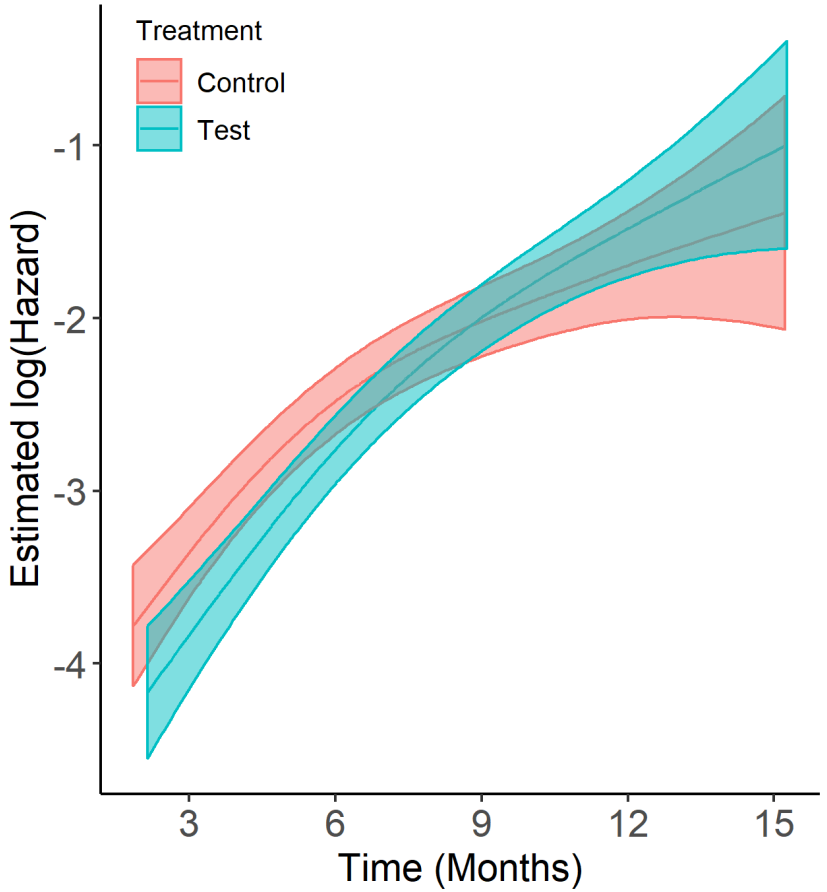
Motivating Example

Treatments: A = test, B = control

simulated data







Grambsch and Therneau (1994) test:
p = 0.015 (evidence of non-PH)

Alternatives to the Logrank Test

- **Weighted logrank tests**
  - Fleming and Harrington (1991) $G^{\rho,\gamma}$ class: weight(t) = $\widehat{S(t)}^{\rho}\left(1-\widehat{S(t)}\right)^{\gamma}$
  - Z1= $G^{0,0}$ (logrank), Z2= $G^{0,1}$ (late), Z3= $G^{1,0}$ (early), Z4= $G^{1,1}$ (middle)
  - MaxCombo test (uses best observed among Z1, Z2, Z3 and Z4)
  - No clinically interpretable estimand

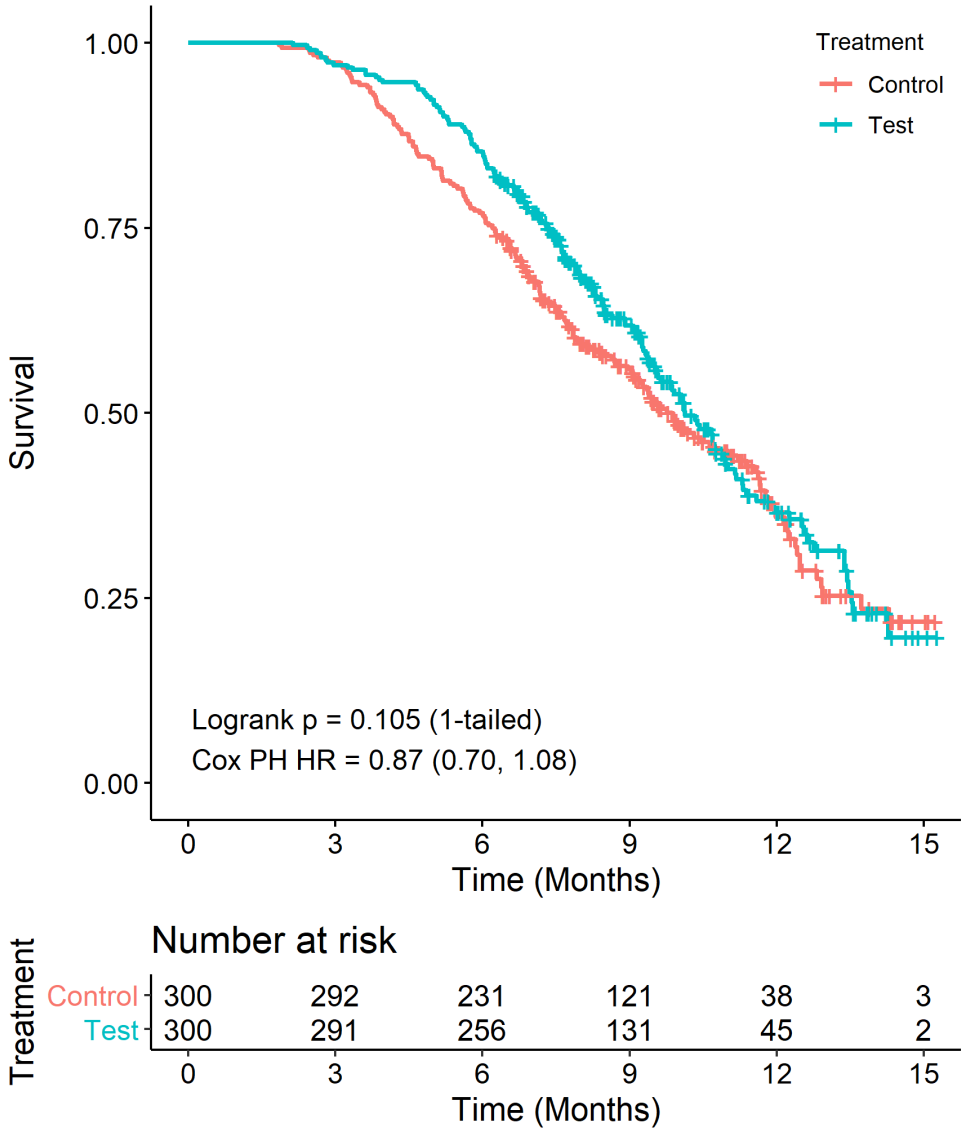
  Roychoudhury et al (2019)

- **Comparison of weighted Kaplan-Meier curves**
  - Special case: Restricted Mean Survival Time (RMST) comparison

    RMST difference: $\delta(\tau) = \int_0^{\tau}[S_A(t) - S_B(t)]dt$

    Royston and Parmar (2011); Tian et al (2014); Uno et al (2014)

Motivating Example (continued)



Analysis Method	1-tailed p-value
Logrank	0.105
MaxCombo	0.057
RMST	0.082

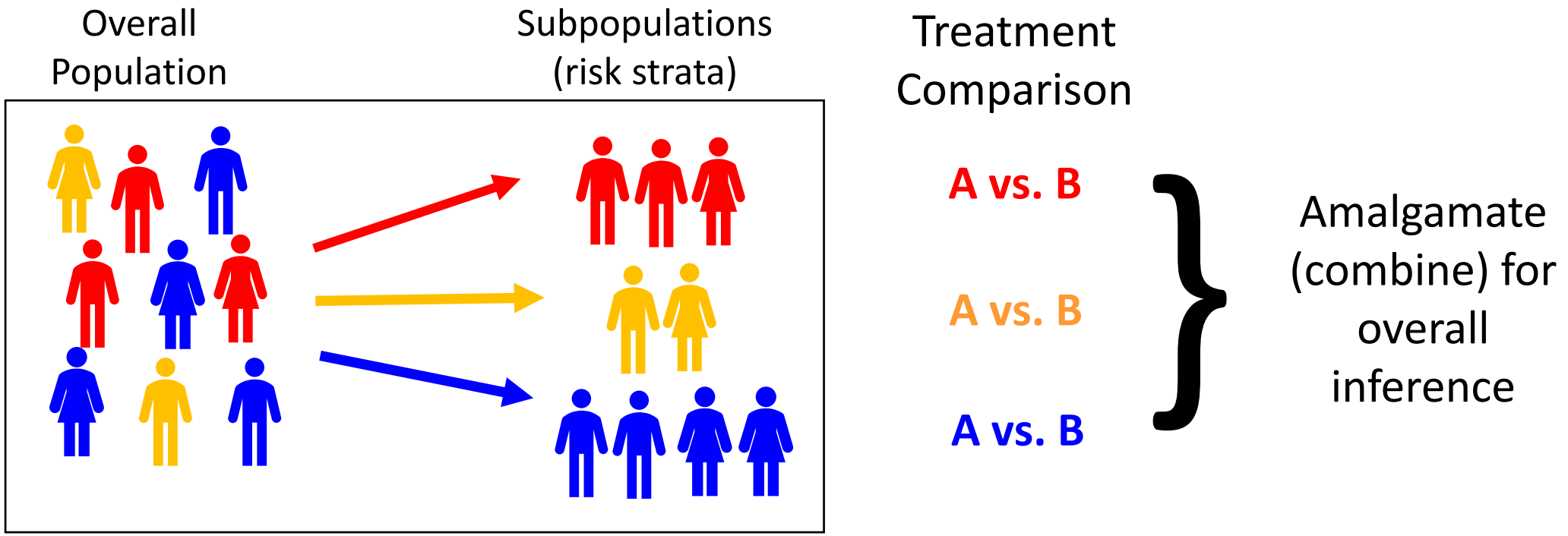
☹

- No treatment effect "signal" using standard approaches
- **Correct conclusion for this dataset?**
- No … shown later

Proposed 5-STAR Approach

Detection of a treatment effect should be easier in a clinically **homogeneous** rather than a **heterogeneous** patient population

5-step Stratified Testing and Amalgamation Routine (5-STAR)



Problem Set-Up: Assumptions, Estimand, Null Hypothesis

Assumptions (within each true risk stratum $i = 1,2, \ldots s$)

- Patients are prognostically homogeneous, i.e., survival times $\{T_{ij}$; risk stratum i, treatment $j\}$ are independently and identically distributed
- $\log(T_{iA})$ is distributed as $\log(T_{iB}) + \Delta_i \Leftrightarrow S_{iA}(t) = S_{iB}\left(e^{\Delta_i} t\right)$

  $\gamma_i = e^{\Delta_i}$ is the **time ratio** in risk stratum *i*

Estimand (overall average treatment effect parameter)

- $\bar{\Delta} = \sum_{i=1}^{s} f_i \Delta_i$ (f_i = proportion of stratum *i* patients in the overall population)

  $\bar{\gamma} = e^{\bar{\Delta}}$ is the true **'average' time ratio**

Null Hypothesis

- $H_0^*: \bigcap_{i=1}^{s} [S_{iA}(t) = S_{iB}(t) \text{ for all } t] \Leftrightarrow H_0^*: \Delta_i = 0$ (i.e., $\gamma_i = 1$) for all i

  Note: $H_0^* \Rightarrow H_0: S_A(t) = S_B(t)$ for all t (on slide 3)

Problem Set-Up: What About Hazard Ratios?

- If $Y_{iA} \sim Y_{iB} + \Delta_i \Leftrightarrow S_{iA}(t) = S_{iB}\left(e^{\Delta_i} t\right)$, will the treatment hazard functions be proportional? Yes, but only in a special case!
- Proportional hazards will hold in risk stratum i if (and only if) T_{iB} ~ **Weibull**
  In this special case, $log[h_{iA}(t)] = log[h_{iB}(t)] + \beta_i \Leftrightarrow S_{iA}(t) = [S_{iB}(t)]^{\theta_i}$
  $\theta_i = e^{\beta_i}$ is the **hazard ratio**
- Supplemental estimand: **'average' hazard ratio** $\bar{\theta} = e^{\bar{\beta}}$, where $\bar{\beta} = \sum_{i=1}^{S} f_i \beta_i$

Waloddi Weibull
1887-1979

Survival times from many RCTs are well-described by a mixture of Weibull distributions

A flexible parametric survival model for fitting time to event data in clinical trials
Jason Liao, Frank Liu. *Pharmaceutical Statistics*. 2019;18:555–567 **[Weibull mixtures]**

Motivating Example (continued): 5-STAR application

Step 1: Pre-specify baseline covariates that might influence survival time

X1, X2, … X50 (includes binary and continuous covariates)

Step 2: Filter out "noise" covariates using Elastic Net Cox regression

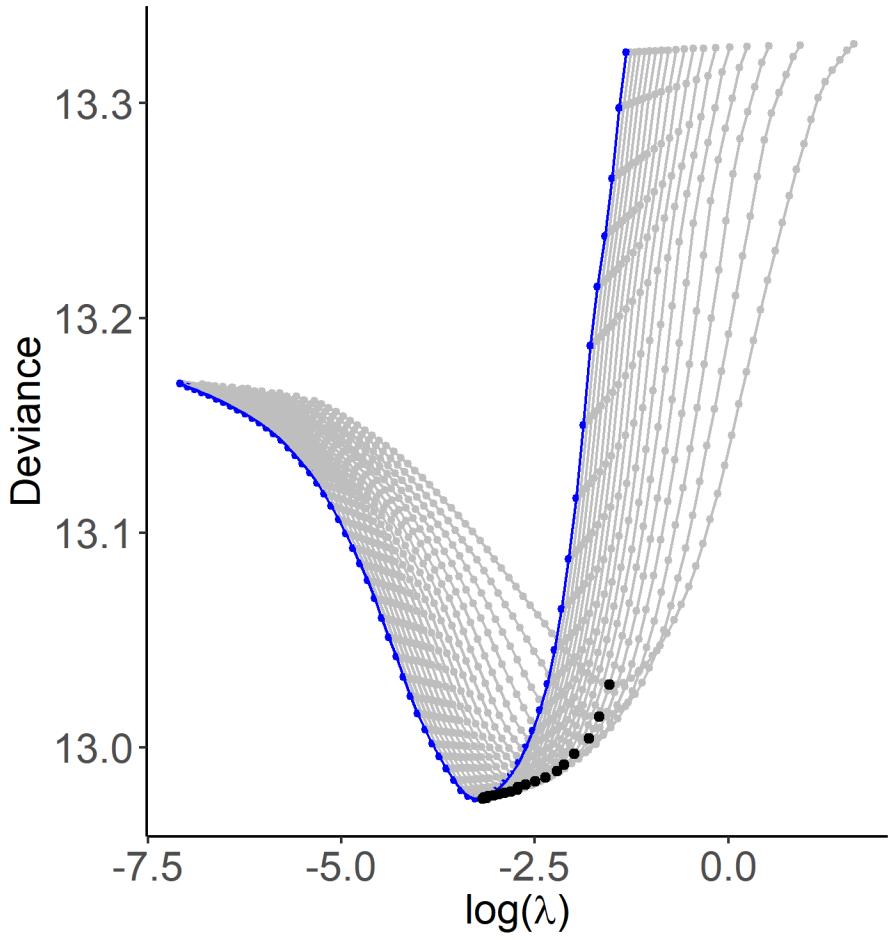
$$\max_{\beta_1 \dots \beta_p} \left\{ \frac{2}{N} logL(y_1 \dots y_N; x_1 \dots x_p, \beta_1 \dots \beta_p) - \lambda \left[\alpha \sum_{k=1}^{p} |\beta_k| + \left(\frac{1-\alpha}{2} \right) \sum_{k=1}^{p} \beta_k^2 \right] \right\}$$

- $L(.)$ = Cox partial likelihood function
- Zou and Hastie (2005), Park and Hastie (2007)
- Pooled survival times (i.e., **without patient-level treatment unblinding**)
- 10-fold cross-validation to optimize λ within a pre-specified α grid

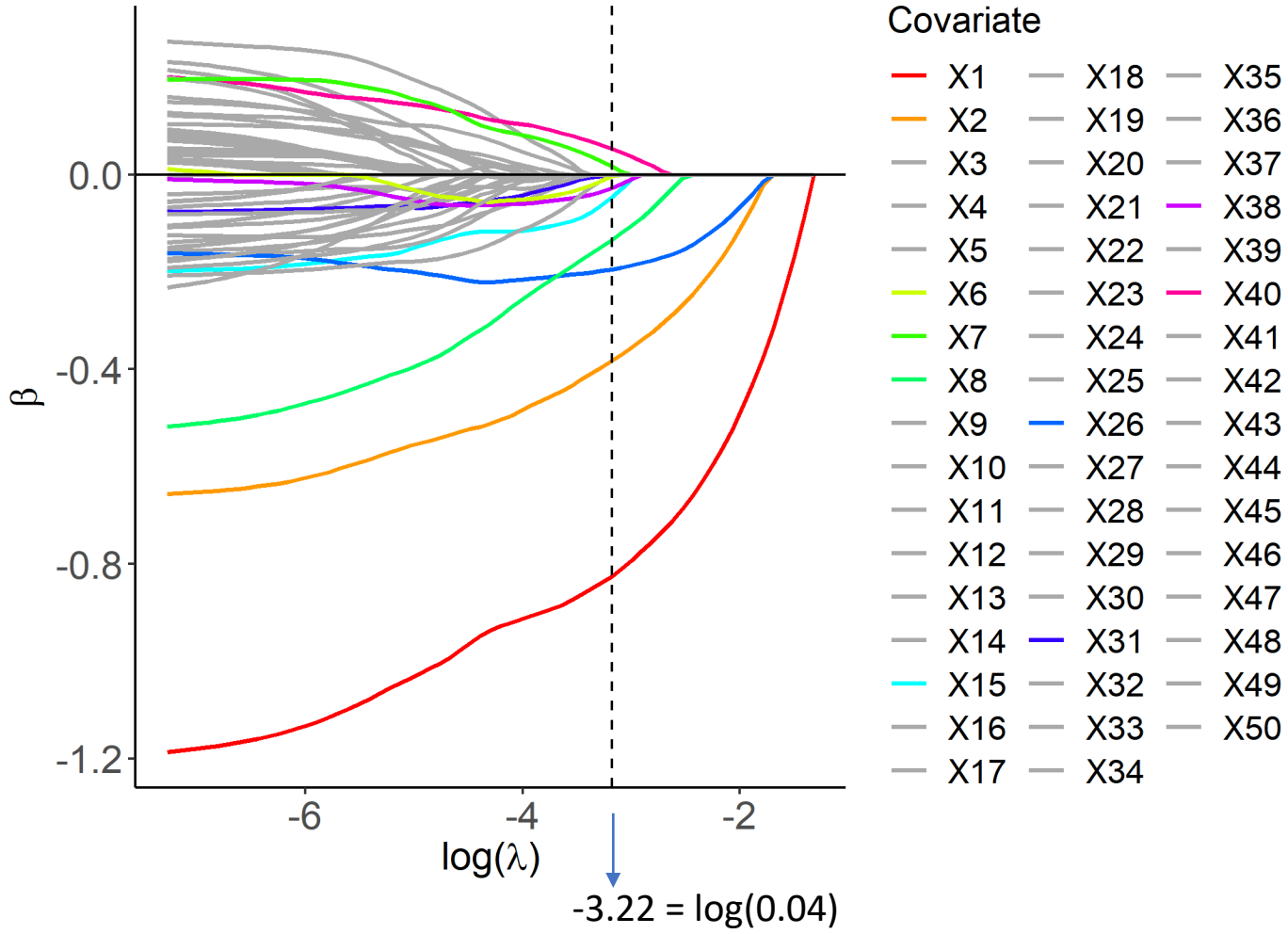
10 covariates kept after elastic net filtering:
X1, X2, X6, X7, X8, X15, X26, X31, X38, X40

Motivating Example (continued): 5-STAR application

Optimization of α and λ



ENET solution path with α_{opt} = 0.95

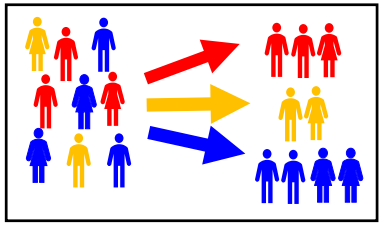


Motivating Example (continued): 5-STAR application

Step 3: Form risk strata using **Conditional Inference Tree** algorithm (Hothorn et al; 2006) **without patient-level treatment unblinding**

3A Form preliminary risk strata

Input: covariates which passed step 2



3B Re-run CTree with *ordered* risk stratum membership from step 3A as a covariate (final risk strata)

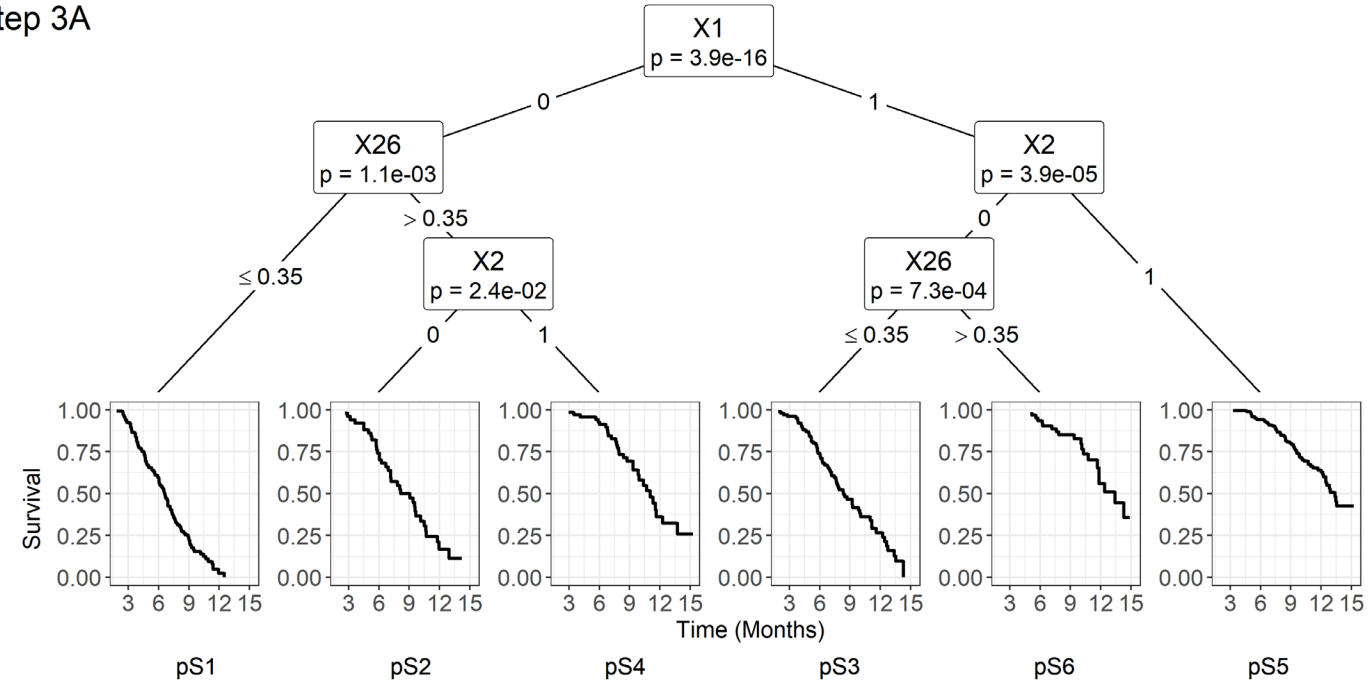
Preliminary Strata:

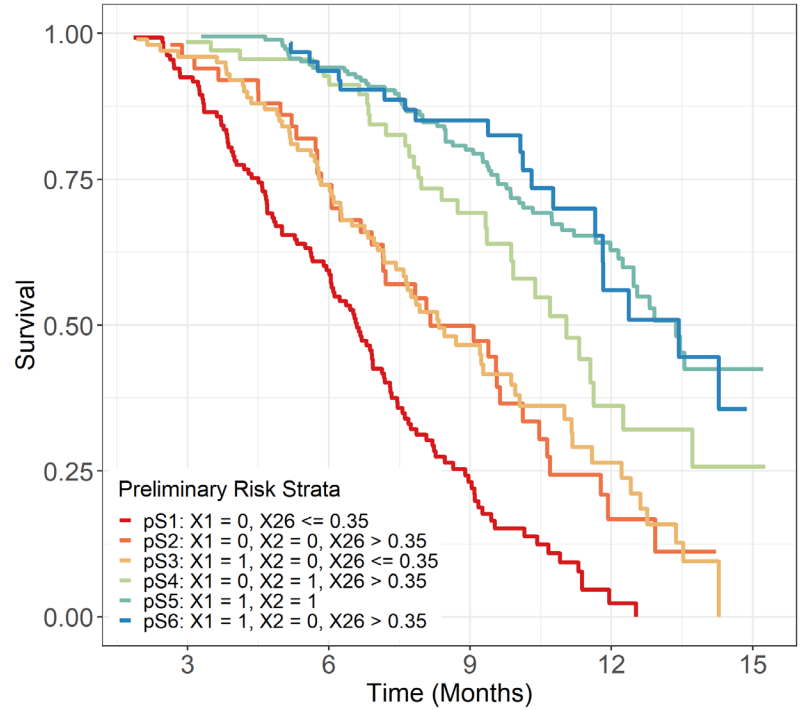
pS1: X1 = 0, X26 ≤ 0.35
pS2: X1 = 0, X2 = 0, X26 > 0.35
pS3: X1 = 1, X2 = 0, X26 ≤ 0.35
pS4: X1 = 0, X2 = 1, X26 > 0.35
pS5: X1 = 1, X2 = 1
pS6: X1 = 1, X2 = 0, X26 > 0.35

Final Strata:

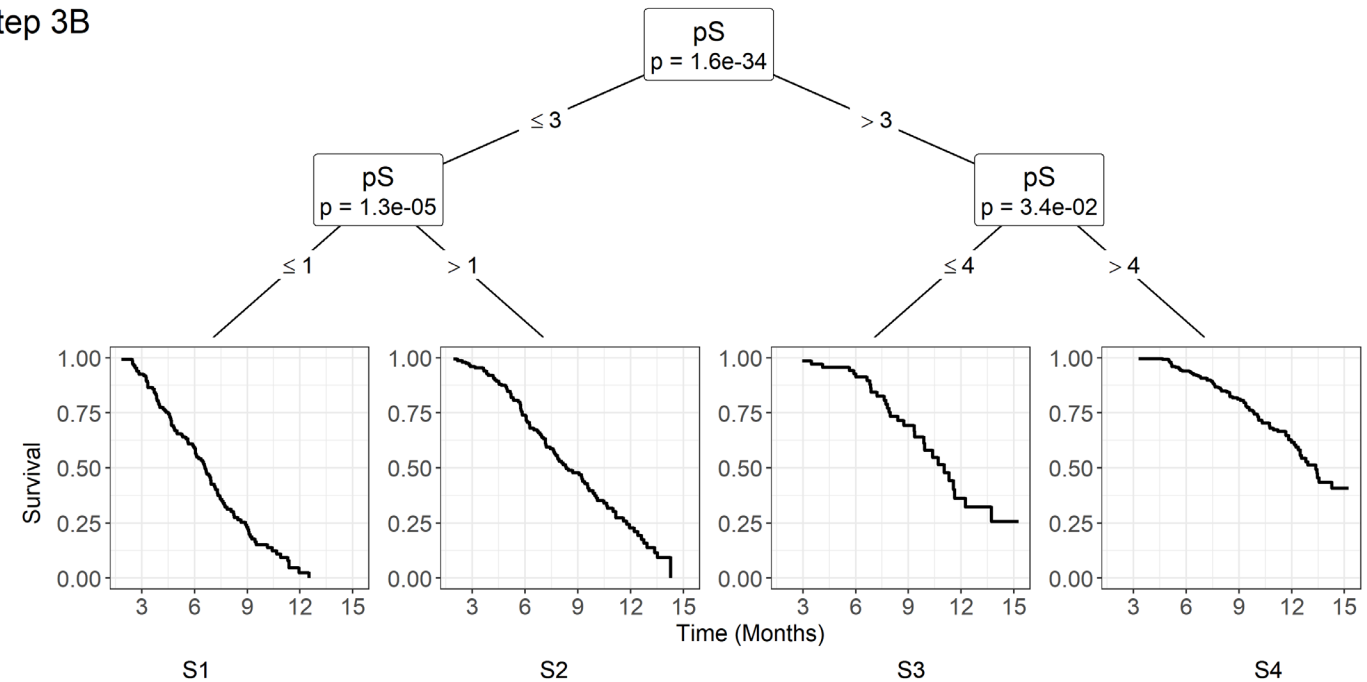
S1: X1 = 0, X26 ≤ 0.35
S2: X2 = 0, {(X1 = 0, X26 > 0.35) or (X1 = 1, X26 ≤ 0.35)}
S3: X1 = 0, X2 = 1, X26 > 0.35
S4: X1 = 1, {(X2 = 1) or (X2 = 0, X26 > 0.35)}

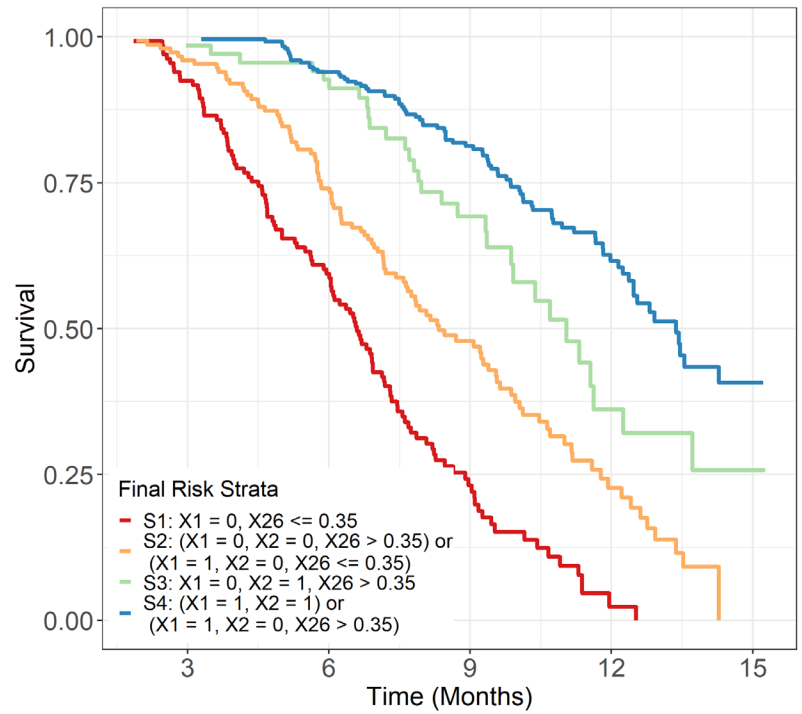
Step 3A





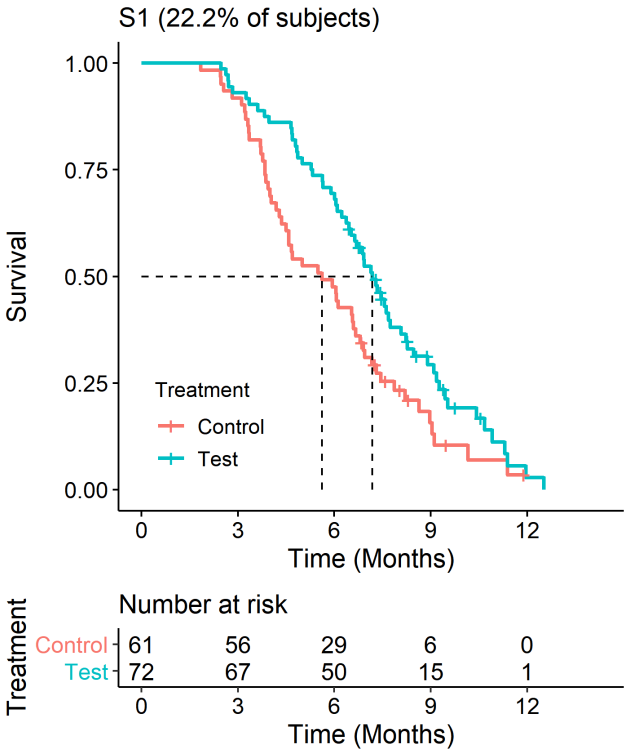
Step 3B

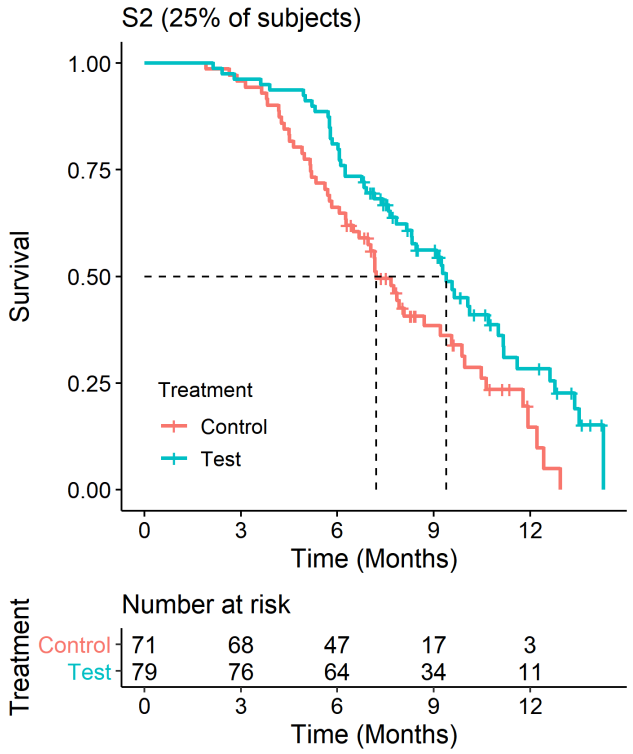


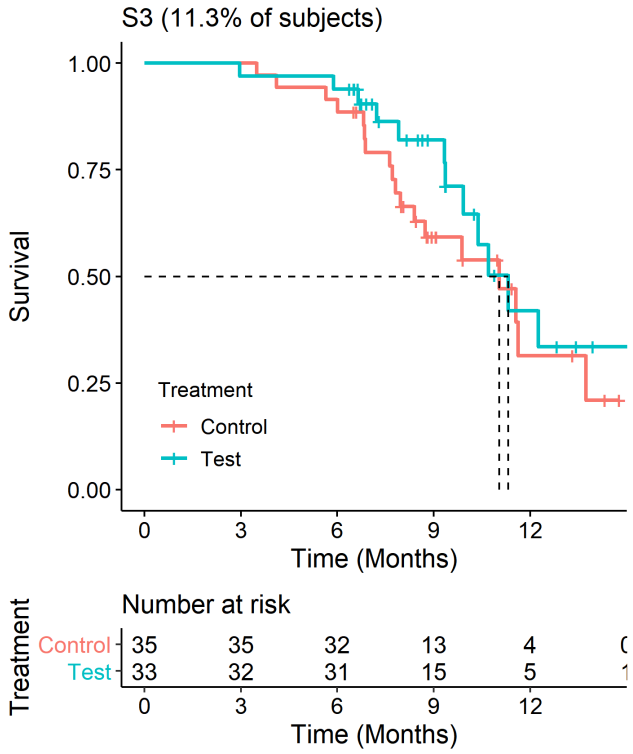


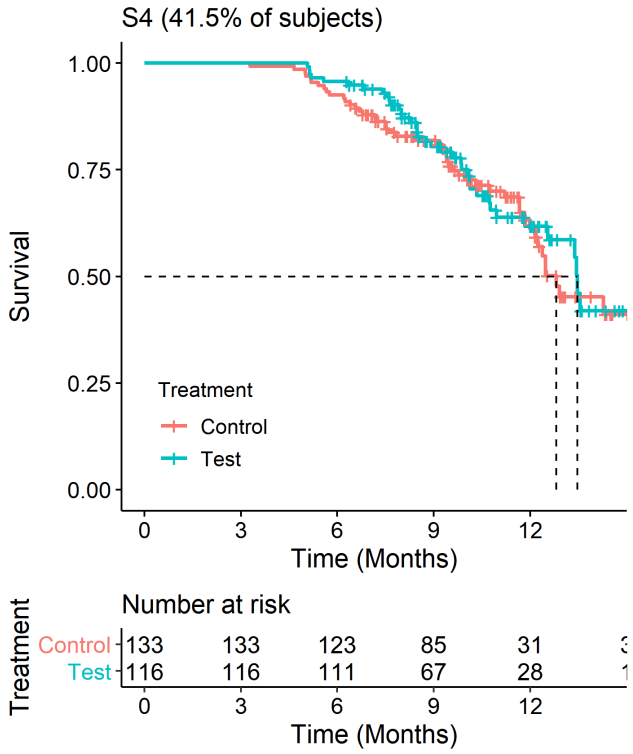
Motivating Example (continued): 5-STAR application

Step 4: Estimate treatment effect within each formed risk stratum









Treatment effect in formed risk stratum q

Primary: $\delta_q = E\left[log\left(T_{qA}\right)\right] - E\left[log\left(T_{qB}\right)\right], \gamma_q = exp\left(\delta_q\right)$ **(Time Ratio; TR)**

Supplemental: $\beta_q = log\left[-logS_{qA}(t)/-logS_{qB}(t)\right]$ *assuming PH*; $\theta_q = exp(q)$ **(Hazard Ratio; HR)**

TR: 1.25 (1.06, 1.47)	**TR**: 1.22 (1.05, 1.43)	**TR**: 1.14 (0.89, 1.47)	**TR**: 1.05 (0.90, 1.22)
HR: 0.62 (0.45, 0.85)	**HR**: 0.59 (0.42, 0.82)	**HR**: 0.71 (0.38, 1.31)	**HR**: 0.90 (0.63, 1.30)

Motivating Example (continued): 5-STAR application

Step 4: primary analysis within formed risk stratum *q* (=1,2, … *c*)

Analysis model: $log(T_{qjk}) = \mu_q + \delta_q I_{qk} + \sigma_q \epsilon_{qjk}$ for formed strat *q*, trt *j*, subj *k*

I_{qk}=1[0] for treatment A[B], $\delta_q = E[log(T_{qAk})] - E[log(T_{qBk})]$

Three parametric fits (T_{ijk} ~ Weibull, log-normal, log-logistic) → model averaging

Obtain $\hat{\delta}_{q,m}$, $V_{q,m}$= $V(\hat{\delta}_{q,m})$, $AIC_{q,m}$ from parametric model fit *m*

$$w_{q,m} = \frac{e^{-0.5AIC_{q,m}}}{\sum_{m=1}^{M} e^{-0.5AIC_{q,m}}}$$

$$\hat{\delta}_q = \sum_{m=1}^{M} w_{q,m}\hat{\delta}_{q,m}$$

$$V_q = \left[\sum_{m=1}^{M} w_{q,m}\sqrt{V_{q,m} + \left(\hat{\delta}_{q,m} - \hat{\delta}_q\right)^2}\right]^2$$

95% CI for Time Ratio (TR) in formed risk stratum *q*: $\exp\{\hat{\delta}_q \mp 1.96\sqrt{V_q}\}$

Motivating Example (continued): 5-STAR application

Step 5: Amalgamate (combine) stratum-level results for overall inference

n_q = number of subjects in formed risk stratum q

$Z_I = \frac{\sum_{q=1}^{c} n_q \widehat{\delta}_q}{\sqrt{\sum_{q=1}^{c} n_q^2 V_q}}$ ~ N(0,1) under H_0^* (asymptotically)

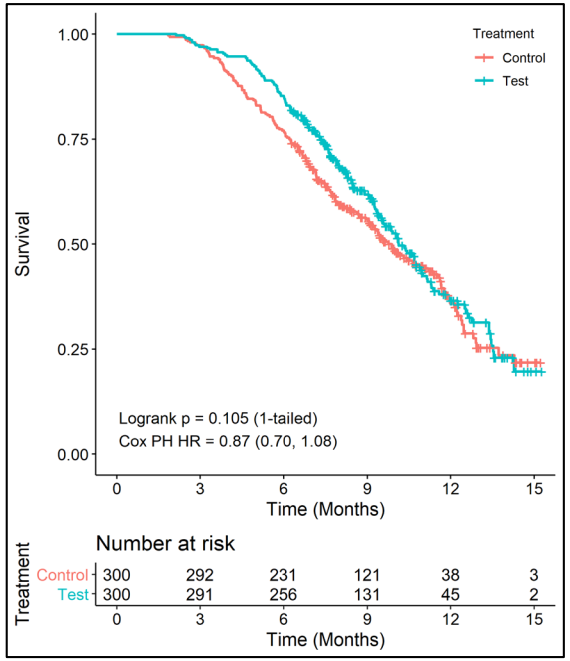
$Z_{II} = \frac{\sum_{q=1}^{c} n_q (\widehat{\delta}_q / \sqrt{V_q})}{\sqrt{\sum_{q=1}^{c} n_q^2}}$ ~ N(0,1) under H_0^* (asymptotically)

$Z_* = max(Z_I, Z_{II})$

Exact distribution: $f(z_*) = 2\phi(z_*)\Phi\left(\frac{1-\rho}{\sqrt{1-\rho^2}} z_*\right)$, $\rho = corr(Z_I, Z_{II})$

Other details in the manuscript

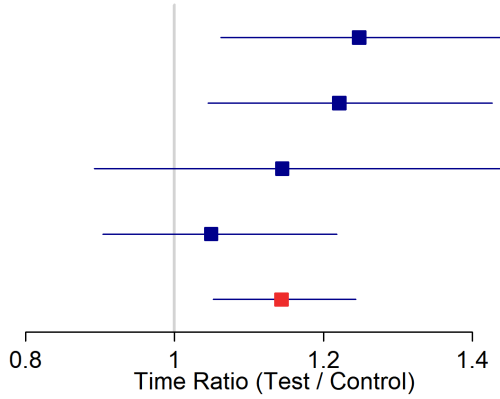
Motivating Example (continued): 5-STAR application



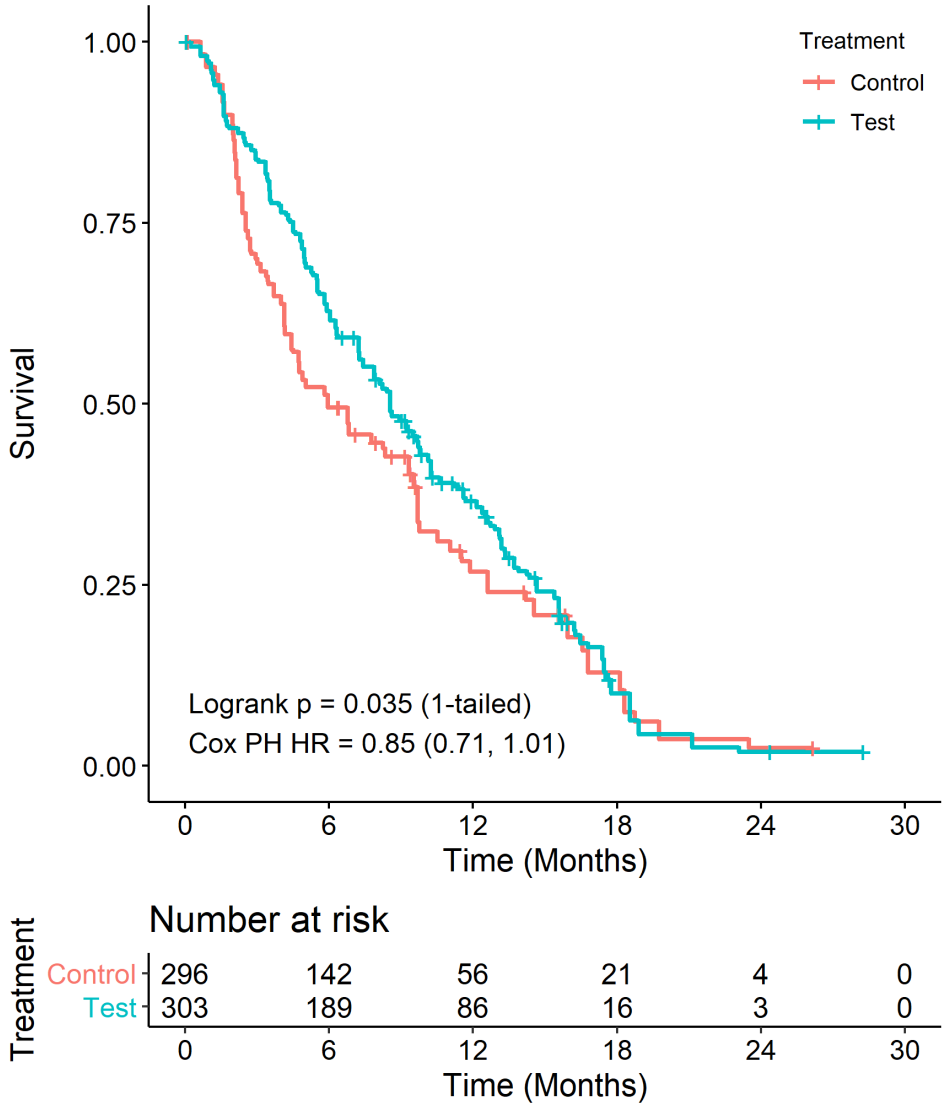
Analysis Method	1-tailed p-value
Logrank	0.105
MaxCombo	0.057
RMST	0.082
5-STAR [TR]	0.001
5-STAR [HR]	0.004

Detailed results for each identified risk stratum

			Primary		Supplemental	
Final Strata	**No. Subjects (%)**	**No. Events (%)**	**Est. TR (95% CI)**	**Pr(TR>1)**	**Est. HR (95% CI)**	**Pr(HR<1)**
S1: X1 = 0, X26 <= 0.35	133 (22.2)	113 (34.2)	1.25 (1.06, 1.47)	0.997	0.62 (0.45, 0.85)	0.993
S2: (X1 = 0, X2 = 0, X26 > 0.35) or (X1 = 1, X2 = 0, X26 <= 0.35)	150 (25.0)	103 (31.2)	1.22 (1.05, 1.43)	0.994	0.59 (0.42, 0.82)	0.995
S3: X1 = 0, X2 = 1, X26 > 0.35	68 (11.3)	30 (9.1)	1.14 (0.89, 1.47)	0.856	0.71 (0.38, 1.31)	0.822
S4: (X1 = 1, X2 = 1) or (X1 = 1, X2 = 0, X26 > 0.35)	249 (41.5)	84 (25.5)	1.05 (0.90, 1.22)	0.736	0.90 (0.63, 1.30)	0.679
5-STAR Average	**600 (100)**	**330 (100)**	**1.14 (1.05, 1.24)**		**0.72 (0.57, 0.92)**	



Example #2 (oncology, N=599)



Grambsch and Therneau (1994) test:
p = 0.002 (evidence of non-PH)

No pre-specified strata

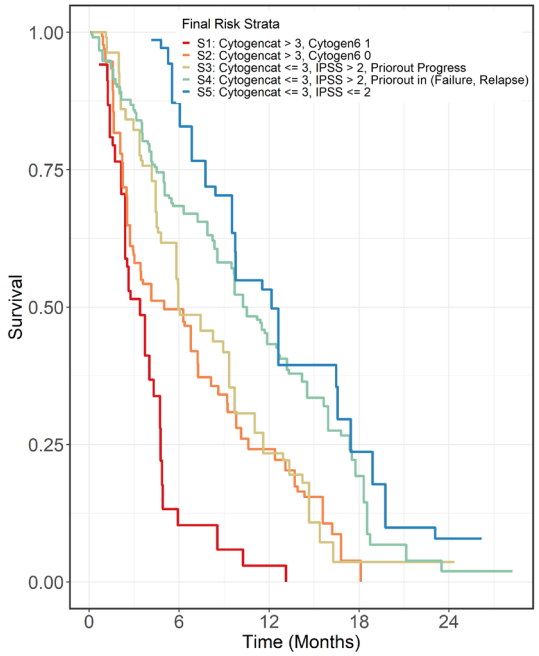
Analysis Method	1-tailed p-value
Logrank	0.035
MaxCombo	0.004
RMST	0.014

5-STAR

Step 1: 14 covariates in candidate set

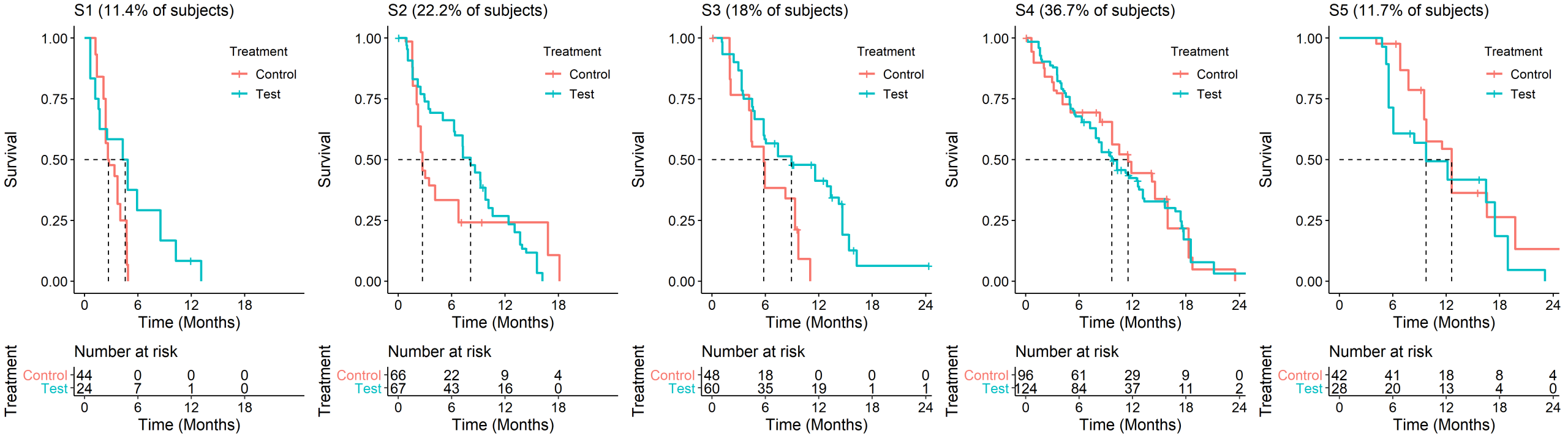
Step 2: 7 covariates advance to step 3

Step 3: 5 risk strata formed based on 4 covariates (below)



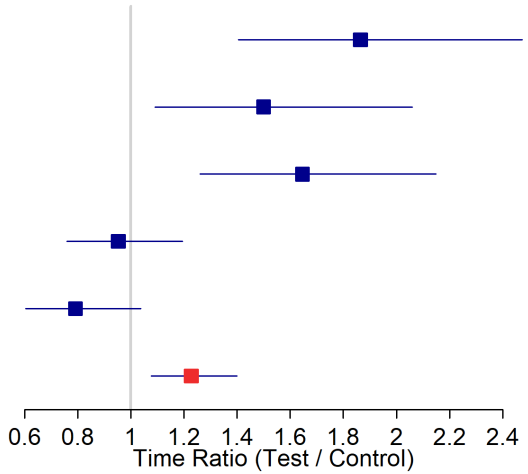
Analysis Method	1-tailed p-value
5-STAR [TR]	0.001
5-STAR [HR]	0.018

Example #2 (continued)

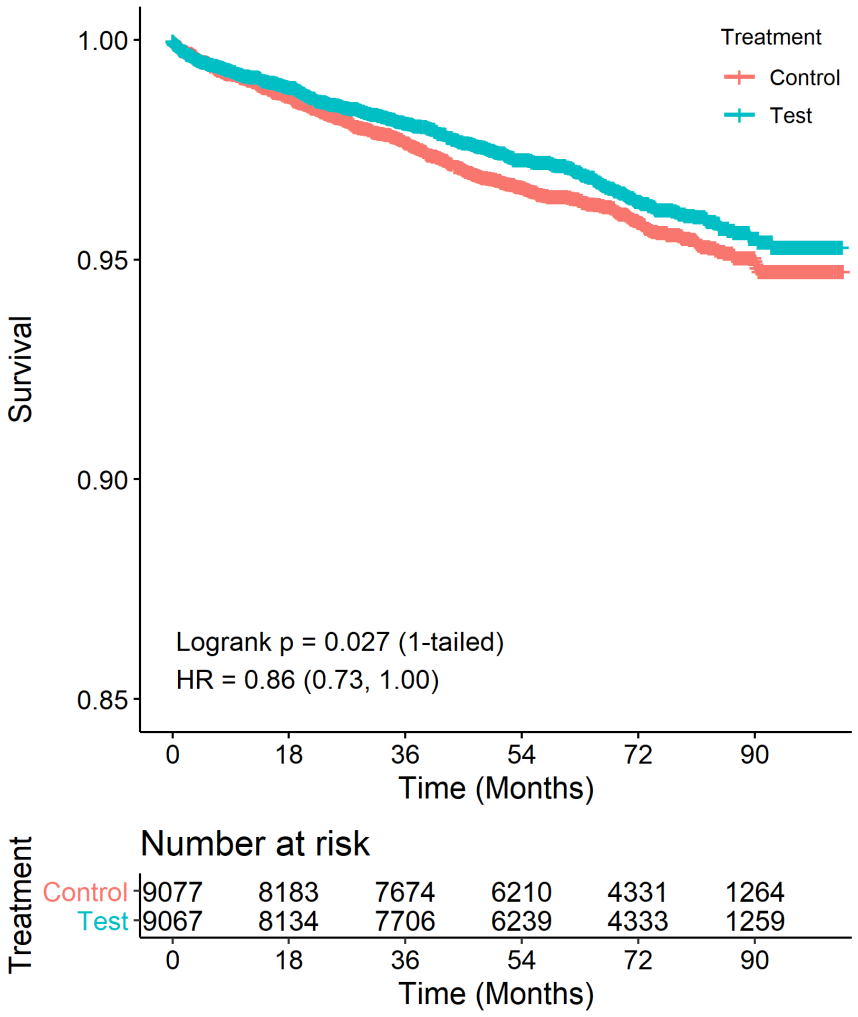


Primary **Supplemental**

Final Strata	No. Subjects (%)	No. Events (%)	Est. TR (95% CI)	Pr(TR>1)	Est. HR (95% CI)	Pr(HR<1)
S1: Cytogencat > 3, Cytogen6 1	68 (11.4)	67 (13.6)	1.86 (1.41, 2.47)	>0.999	0.38 (0.22, 0.64)	0.999
S2: Cytogencat > 3, Cytogen6 0	133 (22.2)	122 (24.7)	1.50 (1.09, 2.06)	0.994	0.94 (0.69, 1.30)	0.617
S3: Cytogencat <= 3, IPSS > 2, Priorout Progress	108 (18.0)	89 (18.0)	1.65 (1.26, 2.15)	>0.999	0.41 (0.28, 0.61)	>0.999
S4: Cytogencat <= 3, IPSS > 2, Priorout in (Failure, Relapse)	220 (36.7)	162 (32.8)	0.95 (0.76, 1.19)	0.336	1.05 (0.80, 1.37)	0.384
S5: Cytogencat <= 3, IPSS <= 2	70 (11.7)	54 (10.9)	0.79 (0.60, 1.04)	0.045	1.59 (1.00, 2.50)	0.049
5-STAR Average	**599 (100)**	**494 (100)**	**1.23 (1.08, 1.40)**		**0.81 (0.66, 0.99)**	

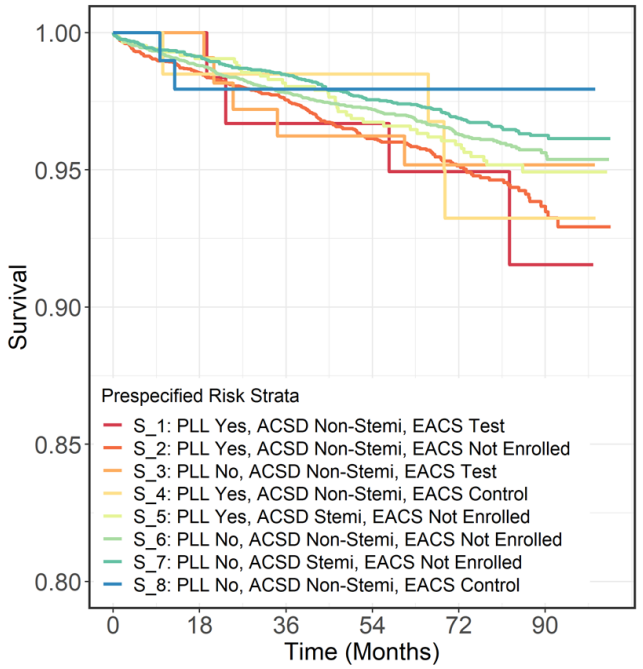


Example #3 (cardiovascular; N=18,144)



Grambsch and Therneau (1994) test:
p = 0.479 (no evidence of non-PH)

Pre-specified strata



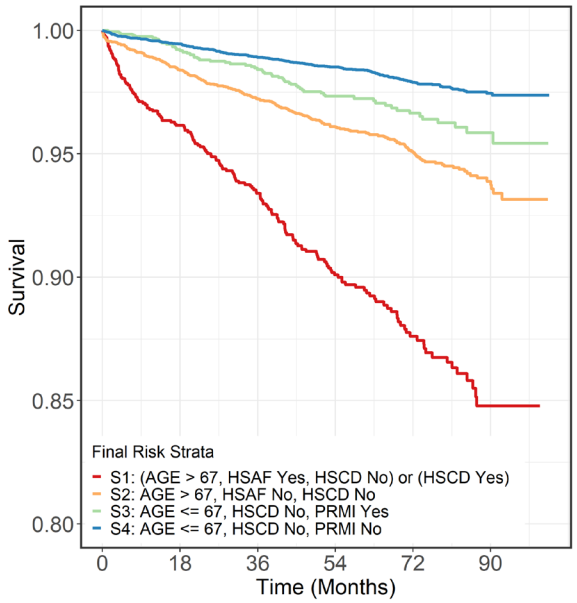
Analysis Method	1-tailed p-value
Logrank	0.027
Stratified logrank	0.026
MaxCombo	0.039
RMST	0.024

5-STAR

Step 1: 46 covariates in candidate set

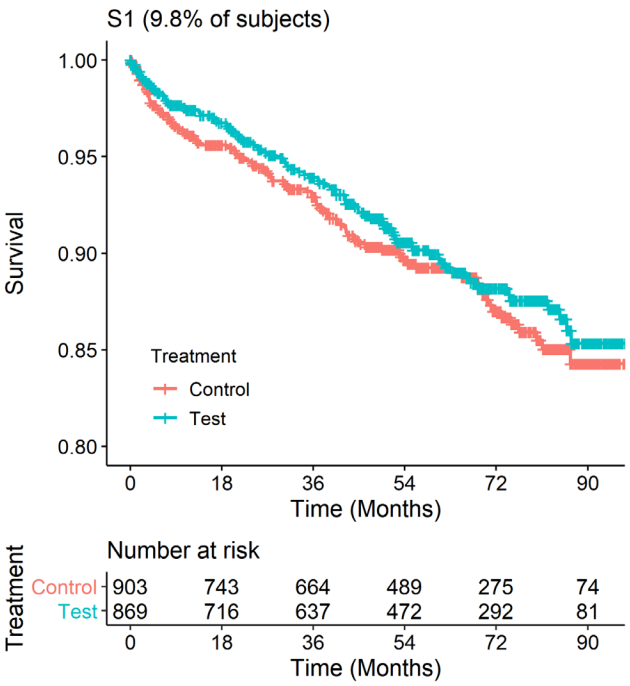
Step 2: 21 covariates advance to step 3

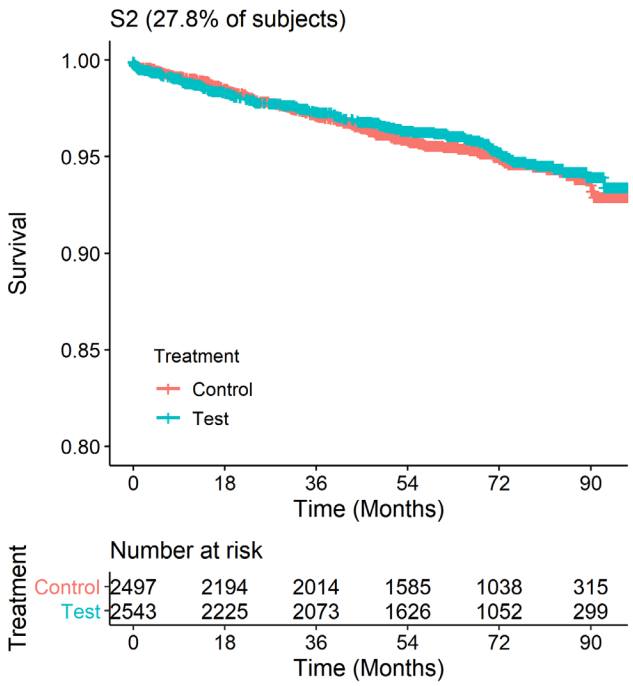
Step 3: 4 risk strata formed based on 4 covariates: age (≤/> 67 yrs), HSAF (yes/no), HSCD (yes/no), PRMI (yes/no)

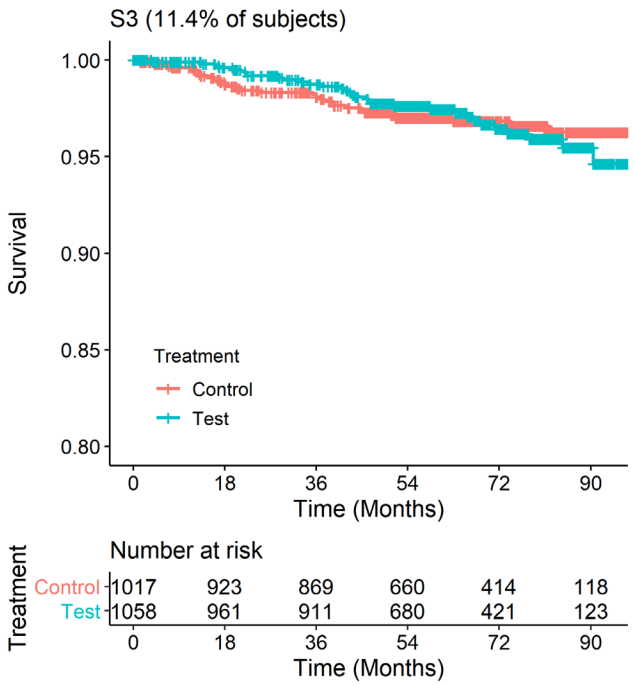


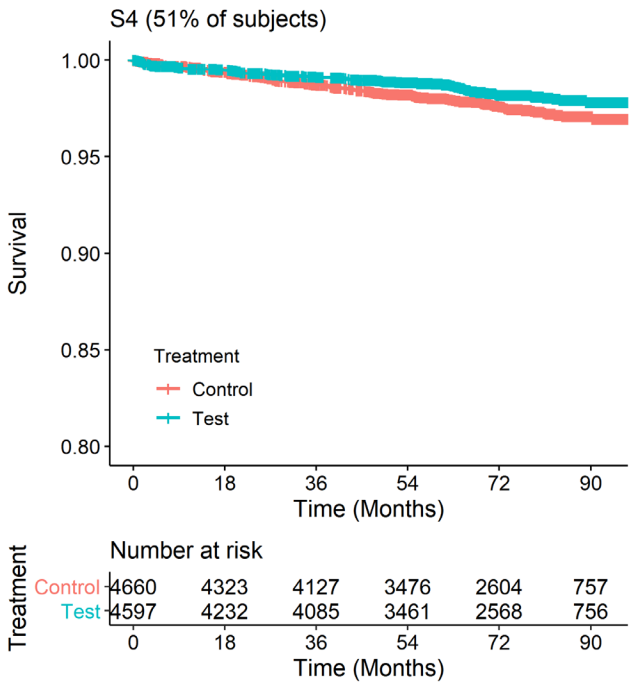
Analysis Method	1-tailed p-value
5-STAR [TR]	0.010
5-STAR [HR]	0.011

Example #3 (continued)

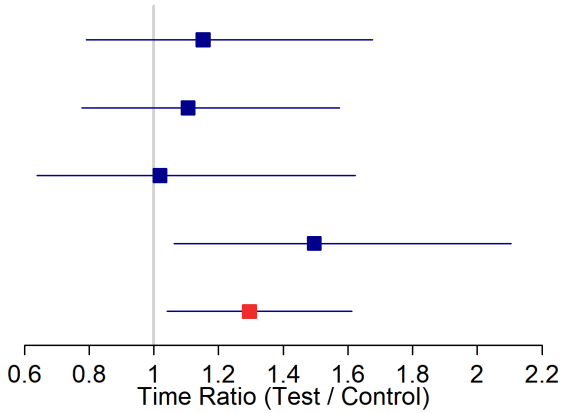




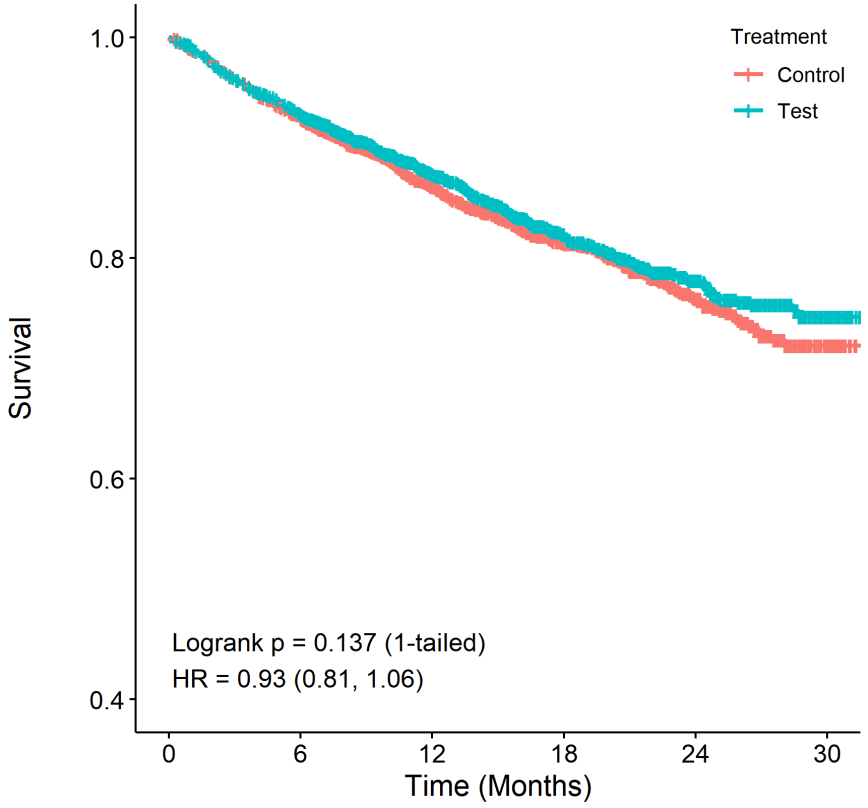




Final Strata	No. Subjects (%)	No. Events (%)	Est. TR (95% CI)	Pr(TR>1)	Est. HR (95% CI)	Pr(HR<1)
			Primary		**Supplemental**	
S1: (AGE > 67, HSAF Yes, HSCD No) or (HSCD Yes)	1772 (9.8)	178 (27.8)	1.15 (0.79, 1.68)	0.770	0.90 (0.70, 1.15)	0.770
S2: AGE > 67, HSAF No, HSCD No	5040 (27.8)	220 (34.3)	1.11 (0.78, 1.57)	0.711	0.93 (0.74, 1.16)	0.710
S3: AGE <= 67, HSCD No, PRMI Yes	2075 (11.4)	62 (9.7)	1.02 (0.64, 1.62)	0.531	1.03 (0.68, 1.56)	0.457
S4: AGE <= 67, HSCD No, PRMI No	9257 (51.0)	181 (28.2)	1.50 (1.06, 2.10)	0.990	0.70 (0.55, 0.90)	0.991
5-STAR Average	**18144 (100)**	**641 (100)**	**1.30 (1.04, 1.61)**		**0.81 (0.67, 0.97)**	

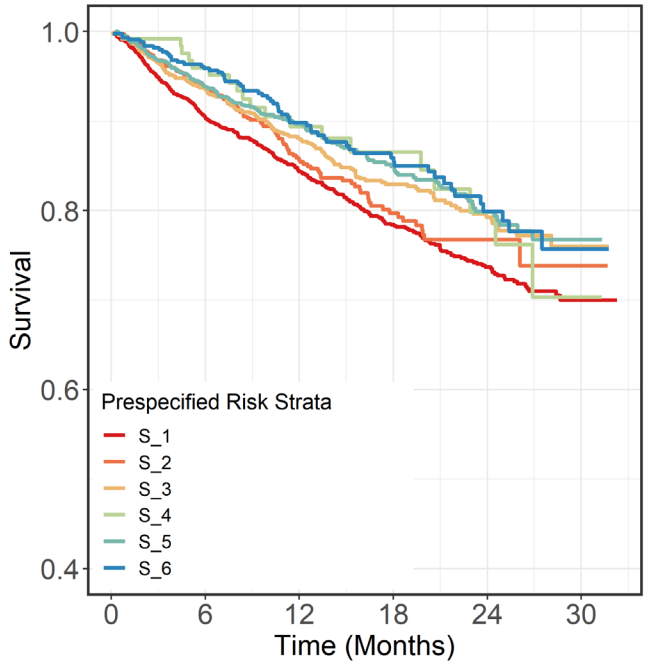


Example #4 (real data)



Grambsch and Therneau (1994) test:
p = 0.667 (no evidence of non-PH)

Pre-specified strata



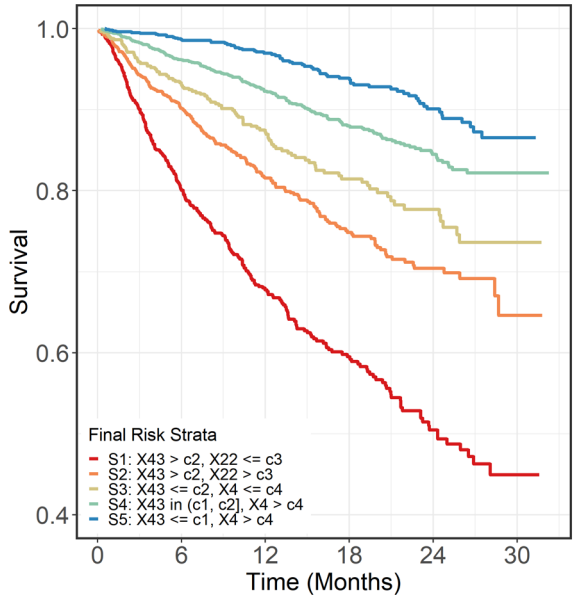
Analysis Method	1-tailed p-value
Logrank	0.137
Stratified logrank	0.134
MaxCombo	0.182
RMST	0.137

5-STAR

Step 1: 65 covariates in candidate set

Step 2: 27 covariates advance to step 3

Step 3: 5 risk strata formed based on 3 covariates: X43 (<=c1/c1-c2/>c2), X22 (<=c3/>c3), X4 (<=c4/>c4)



Analysis Method	1-tailed p-value
5-STAR [TR]	0.023
5-STAR [HR]	0.018

Simulation Study

N=300/trt, target number of events = 330 Truth: 4 risk strata based on (X1, X2, X26>0.4)*					Simulation Scenarios							
					Null HR=1,TR=1		**Alt 1** Equal HRs		**Alt 2** Increasing HRs		**Alt 3** Decreasing HRs	
Risk Stratum	**X1**	**X2**	**X26**	**Median surv. (trt B; control)**	**HR**	TR	**HR**	TR	**HR**	TR	**HR**	TR
S1 (highest risk)	0 0	0 1	≤ 0.4 ≤ 0.4	6.0 months	**1**	1	**0.70**	1.15	**0.42**	1.42	**0.95**	1.02
S2	0 1	0 0	> 0.4 ≤ 0.4	8.4 months	**1**	1	**0.70**	1.13	**0.70**	1.13	**0.86**	1.05
S3	0 1	1 1	> 0.4 ≤ 0.4	10.8 months	**1**	1	**0.70**	1.11	**0.86**	1.04	**0.70**	1.11
S4 (lowest risk)	1 1	0 1	> 0.4 > 0.4	13.2 months	**1**	1	**0.70**	1.09	**0.95**	1.01	**0.42**	1.24

$\bar{\beta} = \sum_{i=1}^{S} f_i \beta_i = \log(0.7)$ in scenarios 1-3, true stratum-averaged HR = exp($\bar{\beta}$) = 0.7; HR=hazard ratio, TR=time ratio

Prevalence: $f_i = 0.25$ for all strata; * among X1-X50 (|corr|≤ 0.45); Weibull distributions in each trt by stratum cell

Simulation Results

20,000 simulated trials

Analysis Method	Type I Error (target α=2.5%)	Power (%)		
		Alt 1 Equal HRs	**Alt 2** Inc. HRs	**Alt 3** Dec. HRs
Logrank	2.56	71	82	50
Stratified logrank*	2.49	77	90	48
MaxCombo	2.60	67	83	54
RMST	2.51	71	84	48
5-STAR [TR]	2.49	84	93	67
5-STAR [HR]	2.52	84	90	73

* analysis based on 2 (of 3) correct and 1 incorrect stratification factors

TR = time ratio, HR = hazard ratio

Conclusions

- Our proposed **5-STAR** approach for survival analysis in RCTs:
  - Boosts power by separating patients into unbiased risk strata and combining stratum-level treatment comparisons for overall inference
  - Does not require a PH assumption within risk strata or overall
  - Delivers “transparency” in overall inference thru stratum-level results
  - Is a promising alternative to current survival analysis methods
  - Is easy to implement (R package) https://github.com/rmarceauwest/fiveSTAR