Overview: Non-Proportional Hazards and Composite Endpoints

Devan V. Mehrotra* and Rachel Marceau West

Biostatistics and Research Decision Sciences

* devan_mehrotra@merck.com

15th Annual Conference on Statistical Issues in Clinical Trials

University of Pennsylvania

April 17, 2023

Non-Proportional Hazards

Logrank Test

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

  T_j = survival time under treatment j, $S_j(t) = Pr(T_j > t)$

  H_{null}: $S_A(t) = S_B(t)$ for all t
- Logrank test = score test from the Cox proportional hazards (PH) model
- When the hazard functions for A and B are **proportional**
  - [Unstratified] 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**
  - [Unstratified] Logrank test is no longer optimal (potential power loss)
  - The Cox PH model HR estimate can be hard to interpret

Some Alternatives to Logrank Test for Tackling Non-PH

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

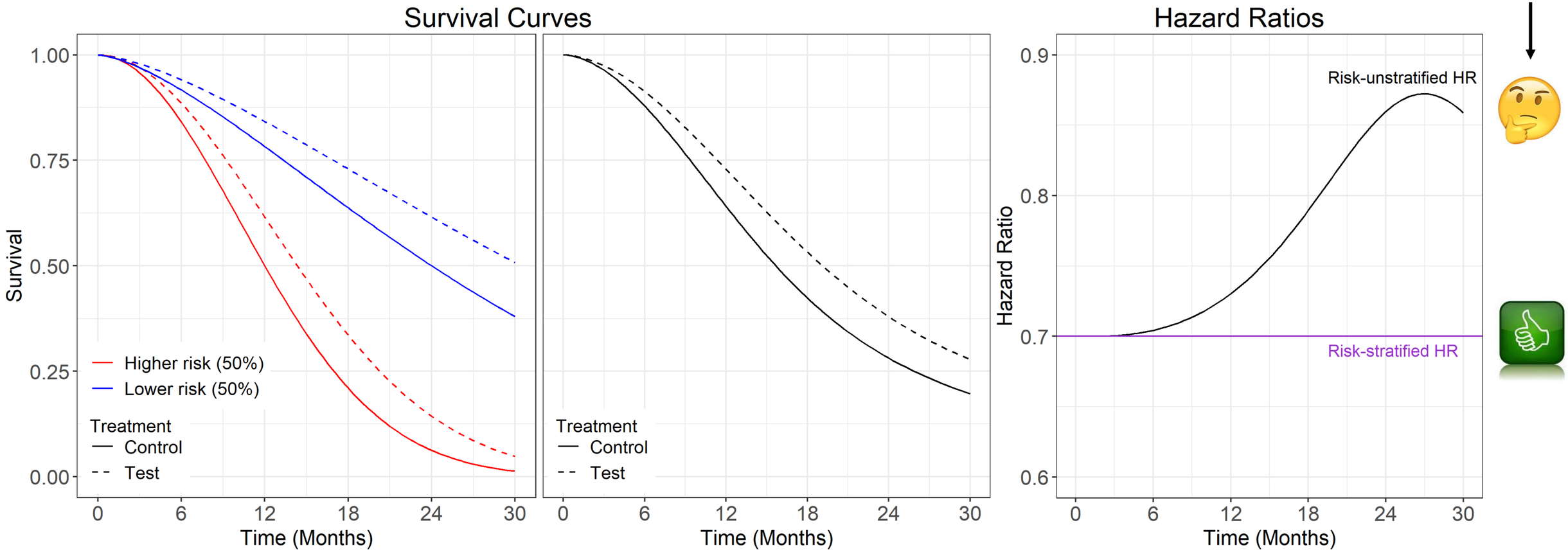
    RMST difference: $\delta(\tau) = \int_0^{\tau}[S_A(t) - S_B(t)]dt$

- These approaches are statistically sound; however, they do not leverage 'structured' prognostic risk heterogeneity commonly anticipated in RCTs
- **No (or inadequate) prognostic risk stratification can create non-PH conditions**

Not Using Prognostic Risk Stratification Can Create Non-PH Conditions

Illustration using a mixture of two Weibull distributions for each treatment

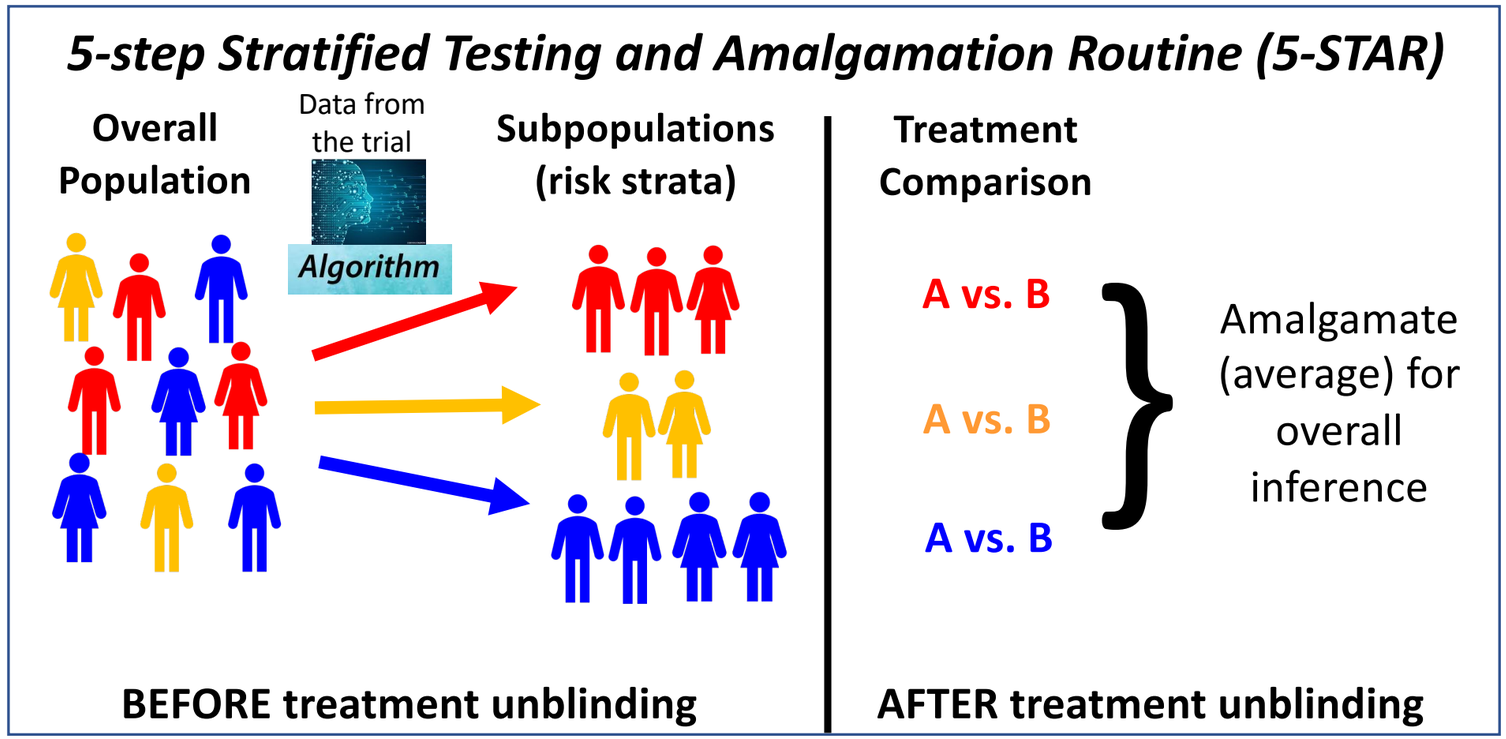
Clinical relevance



True HR = 0.70 (test/control) for lower and higher risk patients

- **Risk-stratified HR** ⇒ averages the HRs for the lower risk and higher risk subpopulations [conceptually]
- **Risk-unstratified HR** ⇒ time-varying value that confounds baseline event risk with treatment effect

Survival Analysis Using Objectively Identified Prognostic Risk Strata



Step 1: Pre-specify baseline covariates that might influence survival time under either treatment
Then, after the trial data have been collected …

BEFORE patient-level treatment unblinding (i.e., based on pooled data across treatment arms)

Step 2: Filter out "noise" covariates using **Elastic Net Cox regression** (Zou and Hastie, 2005)

Step 3: Segment patients into risk strata using **Conditional Inference Tree** (Hothorn et al, 2006)

AFTER patient-level treatment unblinding

Step 4: Estimate treatment effect within each formed risk stratum

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

Important: details for each step in 5-STAR must be ***pre-specified***

Details: Mehrotra DV and Marceau West R, Statistics in Medicine, 39, 4724-4744 (2020)

Simulation Study

	N=300/trt, target number of events = 330 Truth: 4 risk strata based on (X1, X2, X26>0.4)*					Set-Up: PH within each true risk stratum but not overall True Hazard Ratios (HRs) ↓			
	Risk Stratum	**X1**	**X2**	**X26**	**Median surv. (trt B; control)**	**Null** HR=1	**Alt 1** Equal HRs	**Alt 2** Increasing HRs	**Alt 3** Decreasing HRs
	S1 (highest risk)	0 0	0 1	≤ 0.4 ≤ 0.4	6.0 months	1	0.70	0.42	0.95
	S2	0 1	0 0	> 0.4 ≤ 0.4	8.4 months	1	0.70	0.70	0.86
	S3	0 1	1 1	> 0.4 ≤ 0.4	10.8 months	1	0.70	0.86	0.70
	S4 (lowest risk)	1 1	0 1	> 0.4 > 0.4	13.2 months	1	0.70	0.95	0.42

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

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

Simulation Results

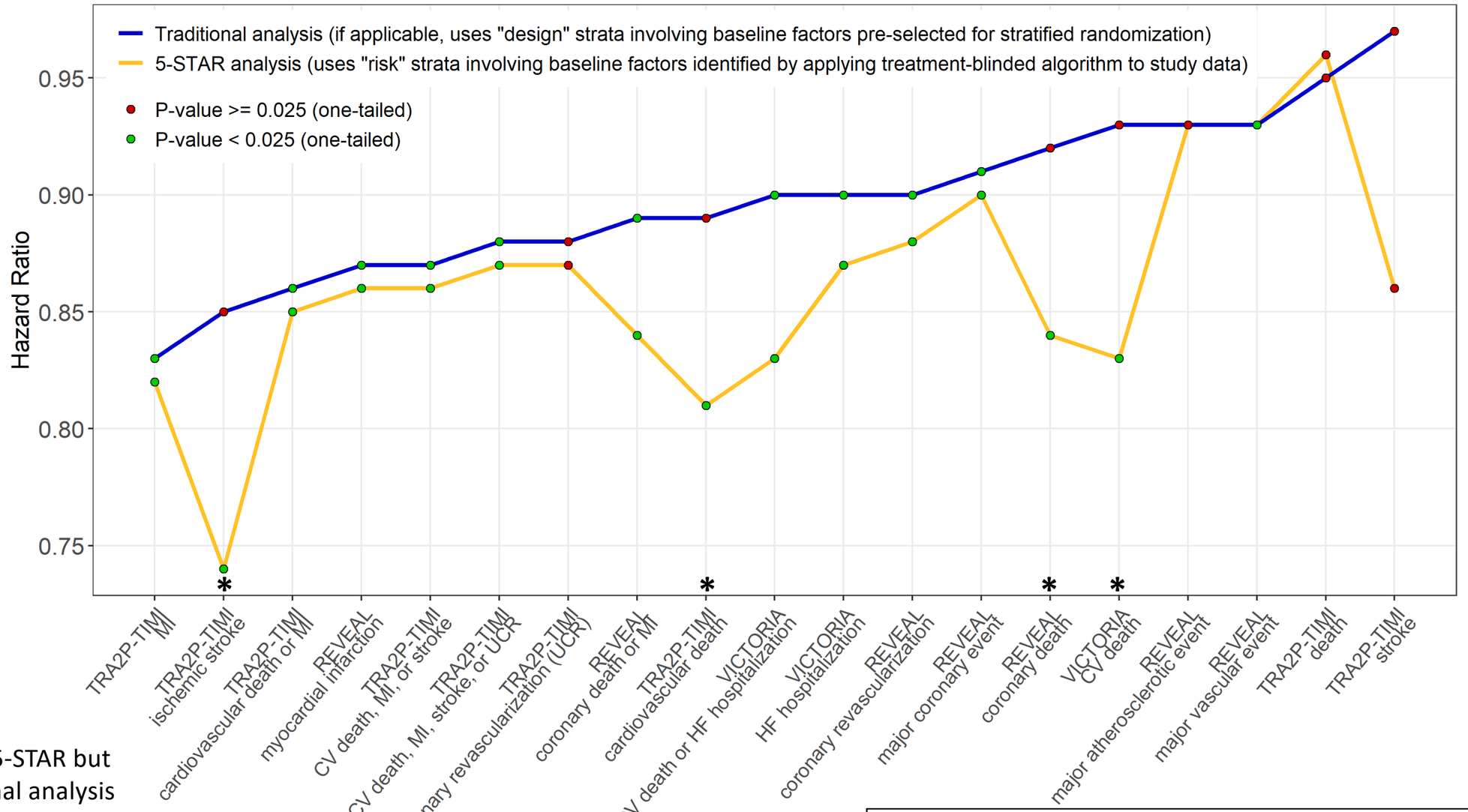
20,000 simulated trials

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	2.52	84	90	73

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

Hazard Ratio Estimates: 19 Real Data Examples

Traditional vs. 5-STAR analysis



* p < 0.025 with 5-STAR but not with traditional analysis

Disclaimer: retrospective analyses for illustration

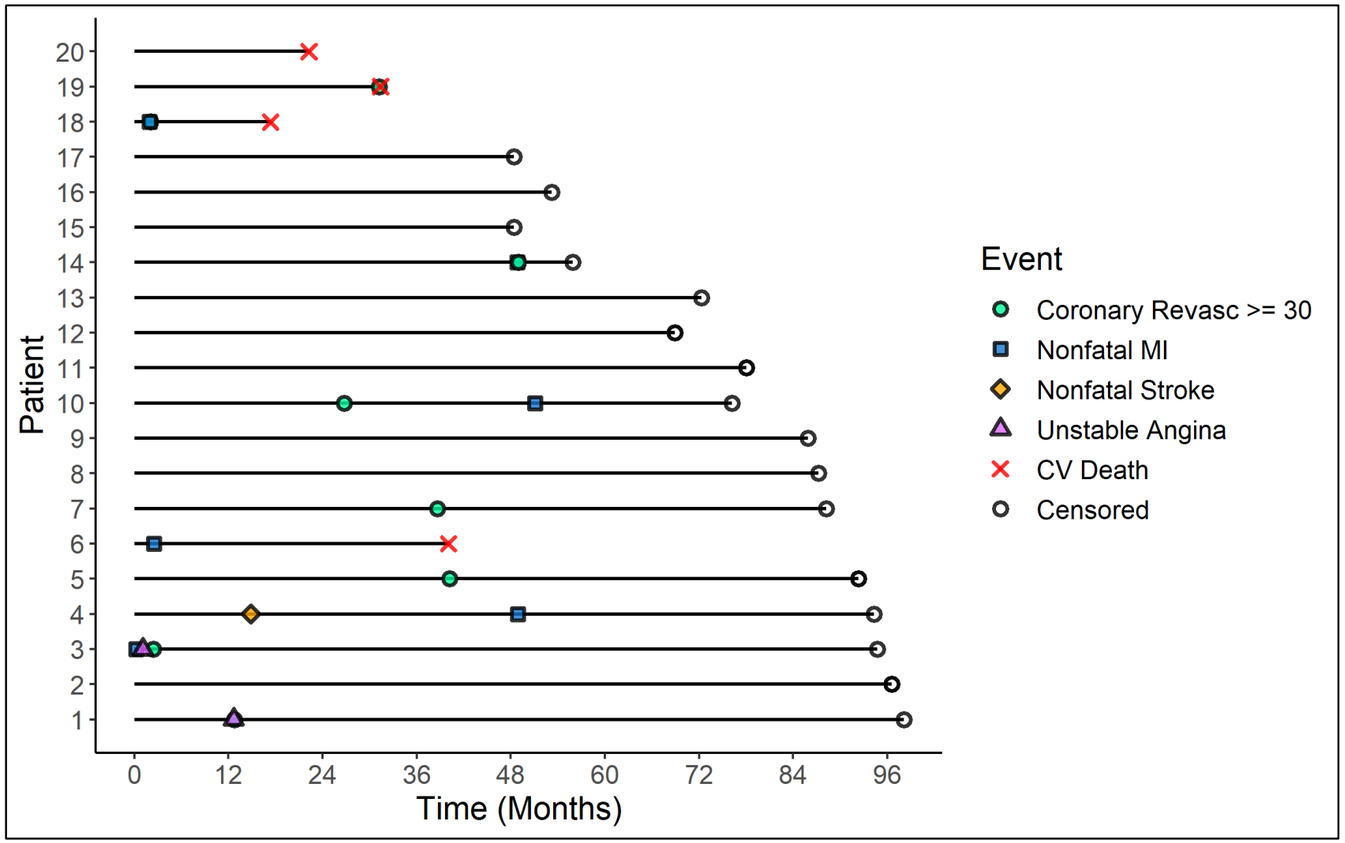
Larger HRs for traditional analysis are [partly] due to adverse impact of no/inadequate risk stratification; seeing "better" HRs using 5-STAR is not a surprise

Composite Endpoints

Motivating Example: IMPROVE-IT Randomized Clinical Trial (N=18,144)

ezetimibe + simvastatin vs. placebo + simvastatin in patients hospitalized for acute coronary syndrome

Longitudinal event profiles for 20 selected patients (for illustration)



Comments

- 5-component **composite endpoint**
- Order of clinical importance
  (1) CV death
  (2) Non-fatal stroke
  (3) Non-fatal MI
  (4) Coronary revasc. ≥ 30 days after random.
  (5) Unstable angina (w/hospitalization)
- Standard analysis: time to first event (TT1E)
  - Ignores event(s) after first event (inefficient)
  - Dominated by "fastest" component; here (4)

Endpoint	Pct. 1st Events	No. Events	HR (95% CI)	1-tailed p-value
CV Death	13%	1075	1.00 (0.89, 1.13)	0.494
Nonfatal MI	32%	2028	0.87 (0.80, 0.95)	0.001
Unstable Angina	4%	304	1.06 (0.85, 1.33)	0.689
Cor. Revasc. >= 30 Days	44%	3483	0.95 (0.89, 1.01)	0.053
Nonfatal Stroke	8%	550	0.80 (0.68, 0.95)	0.005
Traditional Analysis (TT1E)	**100%**	**5314**	**0.94 (0.89, 0.99)**	**0.008**

Total events/Total first events = 7440/5314=**1.40**

Traditional and Alternative Analysis Approaches

Analysis Type	#	Analysis Approach	Output and Reference(s)
Traditional	1	Analysis of time to first event (TT1E)	• HR estimate, CI, p-value
Multiple Events	2	Combine analysis of TT1E, TT2E … assuming neither hazard nor treatment HR change after each event	• HR estimate, CI, p-value • Anderson and Gill (1982)
	3	Combine analysis of TT1E, TT2E … assuming hazard changes after each event but treatment HR is constant	• HR estimate, CI, p-value • Prentice, Williams and Peterson (1981)
Win Ratio	4	Aggregate pairwise subject-level between-treatment comparison of survival times based on sequential order of endpoint importance	• Win Ratio estimate, CI, p-value • Pocock et al (2012)
Combine Individual Component Results	5	Use Cox PH model for each component, combine resulting p-values (equal weights)	• HR estimate, CI, p-value [estimation: invert test] • Brown (1975), Kost & McDermott (2002)
	6	Use Cox PH model for each component, average resulting test statistics (equal weights)	• HR estimate, CI, p-value [estimation: invert test] • Our extension of Stouffer (1949)
	7	Use Cox PH model for each component, average resulting [log] HR estimates (INVAR weights)	• HR estimate, CI, p-value • Wei and Lachin (1984), Lachin and Bebu (2015)
AUC Method	8	Quantify mean cumulative count of events over time using AUC to compare treatments	• AUC ratio estimate, CI, p-value • Claggett et al (2022)

Alternative Analysis Type: Multiple/Recurrent Events

Andersen and Gill 1982; Prentice, Williams, and Peterson 1981

Recurrent events framework extends the Cox PH model to incorporate multiple events per patient (i.e., beyond the first event)

- Patients experiencing a non-fatal event remain in the risk set

Andersen-Gill (AG)

- Patient risk is not impacted by different endpoints or number of events experienced
- Assumes a common hazard over events

Table 1 Data frame for AG model

ID	group	start	stop	status
1	1	0	2	1
1	1	2	3	1
1	1	3	5	1
1	1	5	8	0
2	1	0	10	0
3	2	0	1	1
3	2	1	6	1
3	2	6	10	0

$$coxph(Surv(start, stop, status) \sim group + cluster(ID), data = DataRec)$$

Prentice-Williams-Peterson (PWP)

- Patient risk is stratified by event sequence (1st event, 2nd event, etc.)
- Allows baseline hazard to change with each subsequent event

Table 2 Data frame for the PWP total time approach

ID	group	start	stop	status	enum
1	1	0	2	1	1
1	1	2	3	1	2
1	1	3	5	1	3
1	1	5	8	0	4
2	1	0	10	0	1
3	2	0	1	1	1
3	2	1	6	1	2
3	2	6	10	0	3

$$coxph(Surv(start, stop, status) \sim group + cluster(ID) + strata(enum), data = DataRec)$$

Analysis Method	Est. HR (95% CI)	2-tailed p-value
Andersen-Gill	0.93 (0.89, 0.98)	0.007
Prentice-Williams-Peterson	0.93 (0.89, 0.98)	0.005

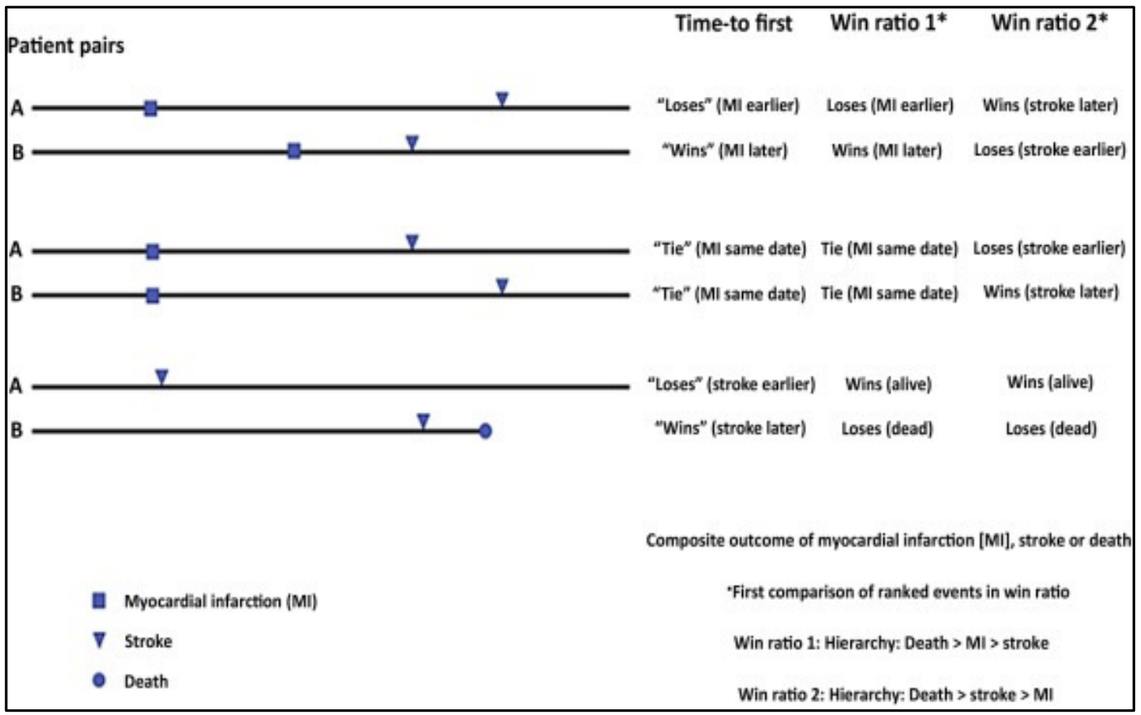
Ozka *et al.* 2018 *BMC Medical Research Methodology*

Alternative Analysis Type: Win Ratio

Pocock et al. 2012, Luo et al. 2015; 2017, Bebu and Lachin 2016, Qui et al. 2017

Hierarchical comparisons consistent with **order of endpoint importance**

- For each pair of subjects from test and control treatment, compare survival times for the most important outcome to determine a "winner" and "loser". Break ties with 2nd most important outcome, and so on



Ferreira *et al.* 2020 *JACC: Heart Failure*

Win ratio

$$WR = \frac{\sum_{k=1}^{K} TW_k}{\sum_{k=1}^{K} TL_k}$$

TW_k: total number of wins for k^{th} outcome

TL_k: total number of losses for k^{th} outcome

Result depends on pre-stated order of clinical importance for component outcomes

Pre-stated Order of Clinical Importance (1 is most important, etc.)	Est. WR (95% CI)	2-tailed p-value
(1) CV Death, (2) Nonfatal Stroke, (3) Nonfatal MI, (4) Coronary Revascularization ≥ 30 days after randomization, (5) Unstable Angina	1.08 (1.02, 1.14)	0.009

Alternative Analysis Type: Combine Individual Component Results

Brown 1975, Kost and McDermott 2002, Poole et al. 2016, Liu and Xie 2019, Stouffer et al. 1949, Strube 1986, Wei and Lachin 1984, Lachin and Bebu 2015

Perform analysis separately within each endpoint of interest and find a smart way to combine results

Combine p-values (Brown 1975, Kost and McDermott 2002, Poole et al. 2016)

$$\mathrm{T} = \sum_{k=1}^{K} -2logp_k$$

p_k: p-value for k^{th} outcome

Overall p-value from scaled Chi-squared distribution, accounting for dependence structure between component outcomes

Combine test statistics (Stouffer et al. 1949, Strube 1986)

$$\mathrm{Z} = \frac{\sum_{k=1}^{K} T_k}{\sqrt{\sum_k Var(T_k) + 2\sum_{i<j} Cov(T_i, Tj)}}$$

T_k: test statistic for k^{th} outcome

Combine log hazard ratios (Wei and Lachin 1984, Lachin and Bebu 2015)

$$Z = \frac{\sum_{k=1}^{K} w_k \hat{\beta}_k}{\sqrt{\hat{V}\left(\sum_{k=1}^{K} w_k \hat{\beta}_k\right)}}$$

$\hat{\beta}_k$: Estimated log hazard ratio for k^{th} outcome

w_k: weight for k^{th} outcome

Analysis Method	Est. HR (95% CI)	2-tailed p-value
Combine p-values (Brown-Kost-McDermott)	0.90 (0.83, 0.97)	0.003
Combine test statistics (Extended Stouffer)	0.93 (0.87, 0.98)	0.019
Combine log HRs (Wei-Lachin w/Inverse Variance Weights)	0.93 (0.88, 0.98)	0.006

Alternative Analysis Type: Mean Cumulative Count of Events (AUC Method)

Claggett et al. 2022

Estimate the area under the mean cumulative count of events curve for each treatment

Interpret as the "*mean total event-free time lost from multiple undesirable outcomes*" over the course of follow-up

Absolute/relative treatment effect is quantified as the difference/ratio of AUCs

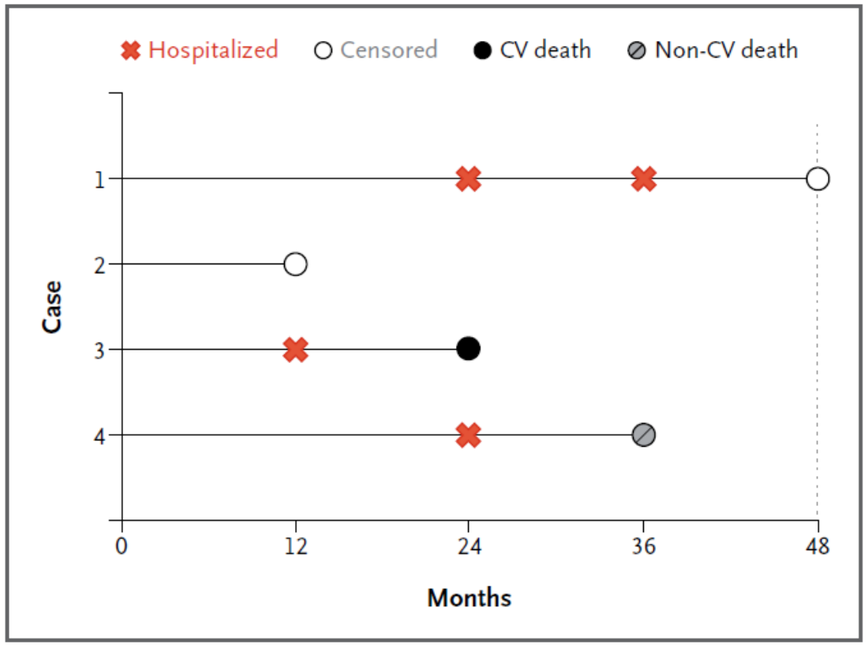


Figure 1. Typical Patterns for the Time-to-Heart-Failure Hospitalization or CV Death from PARAGON-HF.

CV denotes cardiovascular.

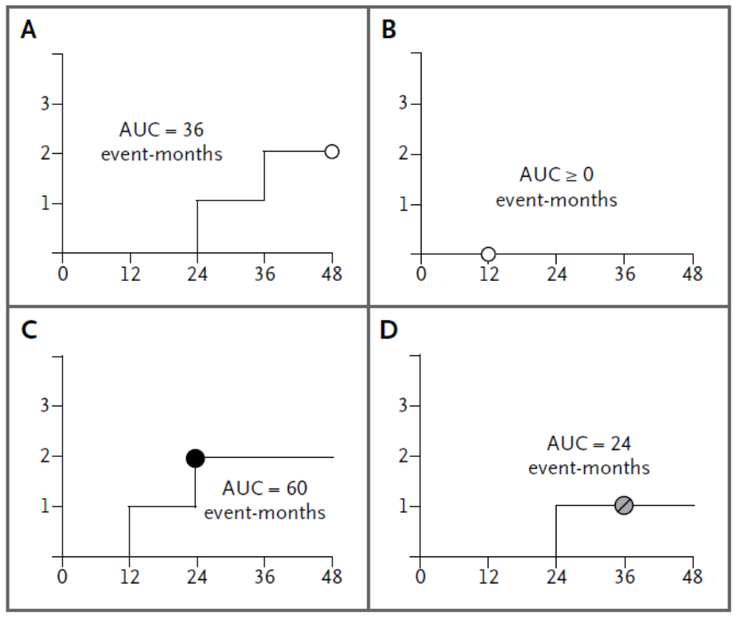


Figure 2. Cumulative Event Curves and AUCs, Corresponding to the Typical Patient Profiles from Figure 1.

Case 1 (Panel A), Case 2 (Panel B), Case 3 (Panel C), and Case 4 (Panel D) are shown. AUC denotes area under the curve. See text for details.

$$A(\tau) = \sum_{i=1}^{n} \sum_{j:X_{ij} \le \tau} \frac{\hat{S}(X_{ij})}{Y(X_{ij})} (\tau - X_{ij}).$$

$A(\tau)$: area under the mean cumulative count curve at time τ

- $\hat{S}(u)$: Kaplan-Meier curve for death from many causes
- $Y(u)$: number of patients still under follow-up at time u
- $\{X_{ij}, j = 1, \ldots, K\}$: times to the K_i events for patient $i, i = 1, \ldots, n$

Analysis Method	2-tailed p-value
Estimate mean cumulative count of events over time by AUC	0.006

Simulation Study - Description

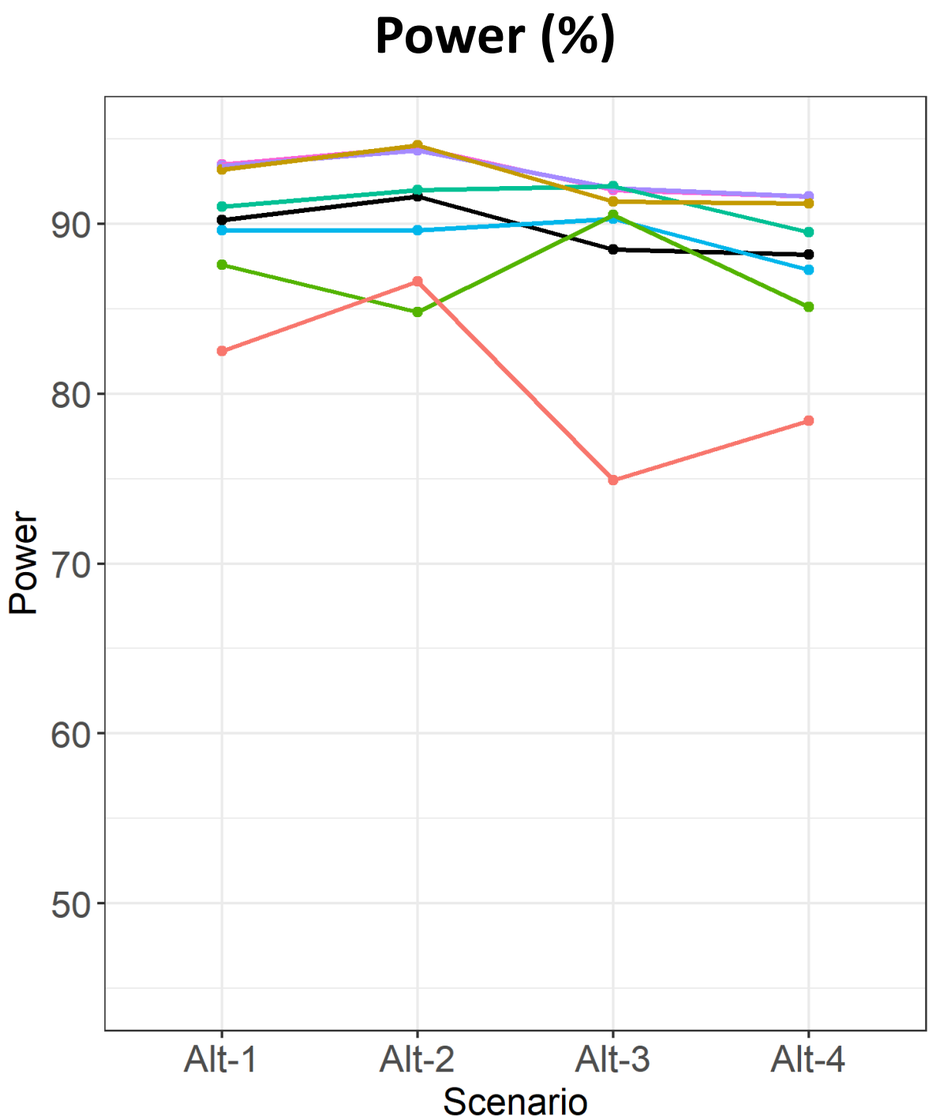
- 1:1 randomization, total N = 4000, 1200 first events
- 90% power to detect HR of 0.80 for primary (TT1E) analysis with 2-tailed $\alpha = 0.01$
- Event times simulated using 5-variate Weibull, with correlations from IMPROVE-IT

Component # First event →	**1** 35%	**2** 32%	**3** 17%	**4** 11%	**5** 5%	**Composite** (1st Event)	
Importance →	3	4	1	2	5		
Null	1	1	1	1	1	1	
Alt-1	0.79	0.79	0.79	0.79	0.79	0.80	
Alt-2	0.78	0.75	0.81	0.84	0.90	0.80	HRs
Alt-3	0.80	0.85	0.76	0.73	0.70	0.80	
Alt-4	0.83	0.78	0.79	0.75	0.85	0.80	

- Total events/total first event = 1461/1200 = **1.22** [conservative; was 1.40 for IMPROVE-IT]
- Median f/up 27 months, first event accrual 14/100 person-yrs, analysis at 45 months

Simulation Results – Power

20,000 simulated trials



Interpretation of Results

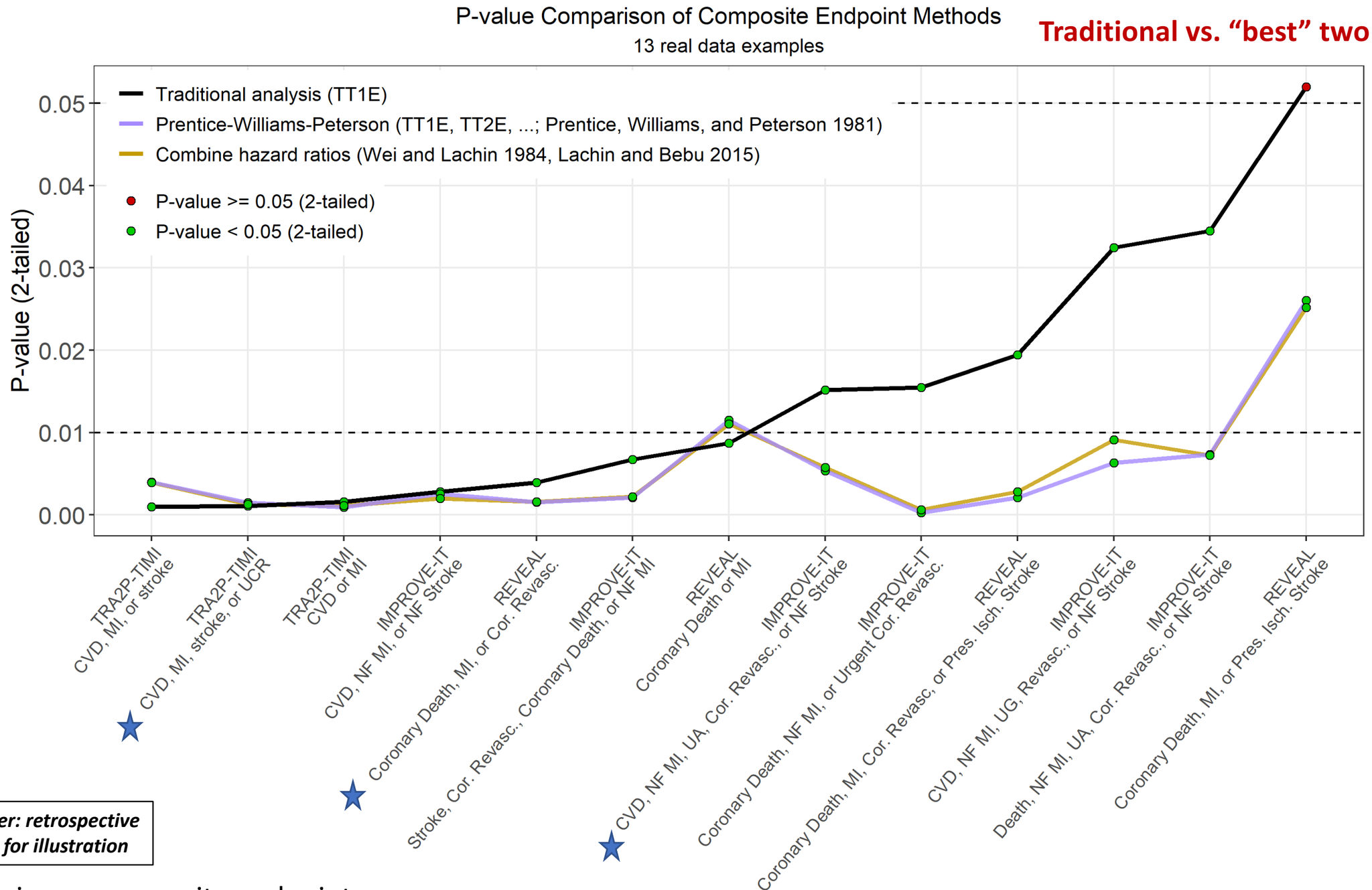
1. Best performers: PWP (multiple/recurrent events) and WL (combine HRs)
2. PWP and WL have ≥ 90% power in every scenario studied
3. To achieve #2 with the traditional (TT1E) analysis would require ~ 10% more events

Method

- Traditional analysis (TT1E)
- Andersen-Gill (TT1E, TT2E, ...; Andersen and Gill 1982)
- Prentice-Williams-Peterson (TT1E, TT2E, ...; Prentice, Williams, and Peterson 1981)
- Win Ratio (Pocock et al. 2012)
- Combine p-values (Brown 1975, Kost and McDermott 2002)
- Combine test statistics (Stouffer 1949)
- Combine hazard ratios (Wei and Lachin 1984, Lachin and Bebu 2015)
- Estimate mean cumulative count of events over time by the area under the curve (AUC) (Clagett et al. 2022)

Not shown: Type I error was well controlled for all methods at $\alpha = 0.01$

Traditional vs. “best” two alternatives



Disclaimer: retrospective analyses for illustration

★ primary composite endpoint

Wrap-Up

Non-Proportional Hazards

- Assessment of non-PH should be aligned with the intended analysis
  - For a stratified analysis, assess non-PH within strata, not overall (i.e., unstratified)
- No or inadequate prognostic risk stratification is often a cause of non-PH
  - Analyses with adequate risk stratification (e.g., 5-STAR) can boost power notably
    - Reporting stratum-level HRs (in addition to their average) is important for interpretation

Composite Endpoints

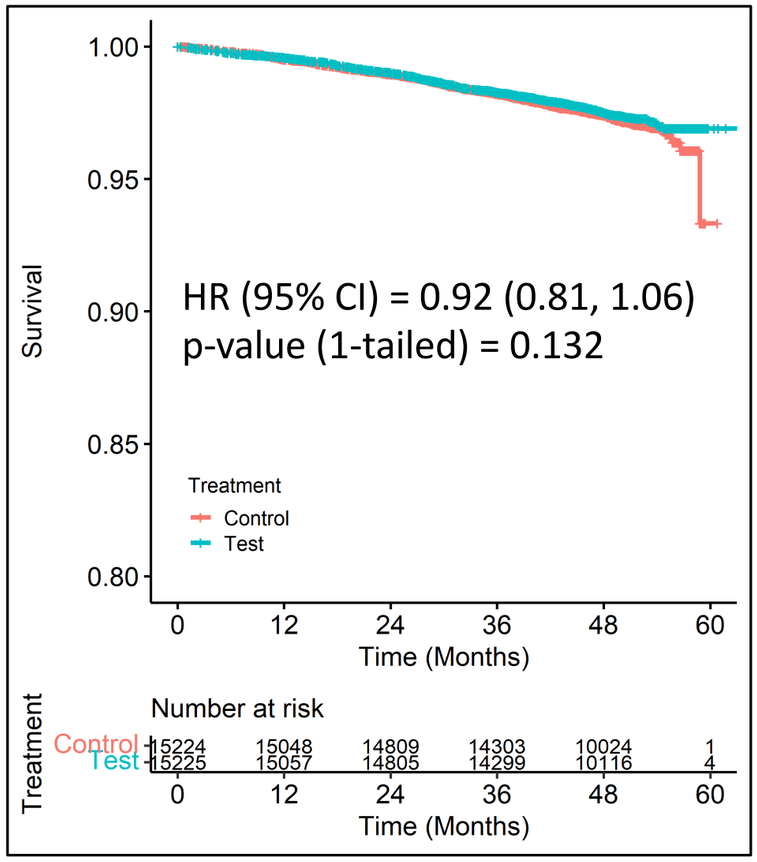
- Methods (such as PWP, WL) that use data beyond the first intra-patient event can improve power relative to the traditional time to first event (TT1E) analysis
  - Sample size reductions of 10-15% are possible in some scenarios
- Other important considerations
  - Interpretation of the reported “treatment effect” [clarity in the estimand]
  - Drug labeling implications and need for upfront regulatory buy-in

Back-Up Slides

REVEAL trial (N=30,449): Coronary Death Endpoint

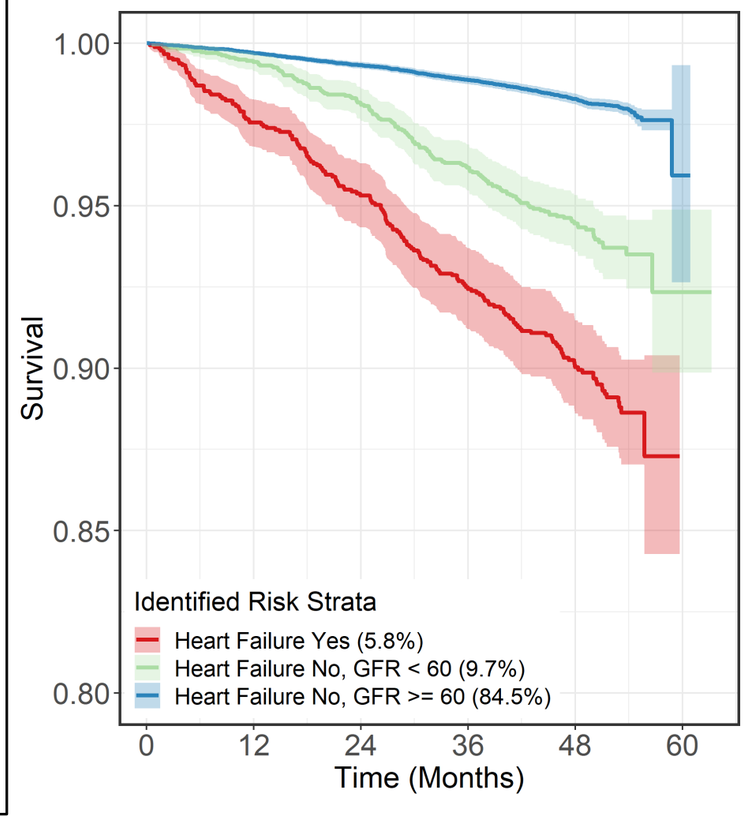
Anacetrapib (test treatment) vs. placebo (control treatment)

Kaplan-Meier curves by treatment



Kaplan-Meier curves by strata

Strata identified using treatment-blinded algorithm (***risk strata***)



No design strata for this trial

	HR Estimate (95% CI)	P-value (1-tailed)
Traditional analysis (unstratified)	0.92 (0.81, 1.06)	0.132
5-STAR analysis (using risk strata)	0.84 (0.71, 0.99)	0.019

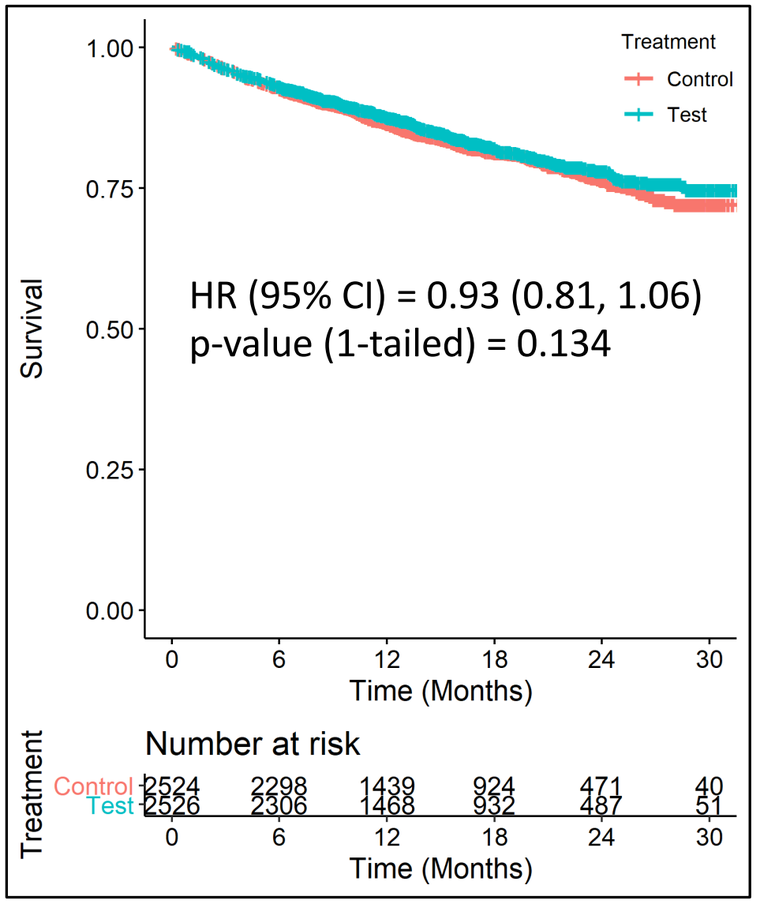
Disclaimer: retrospective analyses for illustration

Disclaimer: retrospective analyses for illustration

VICTORIA trial (N=5,050): CV Death Endpoint

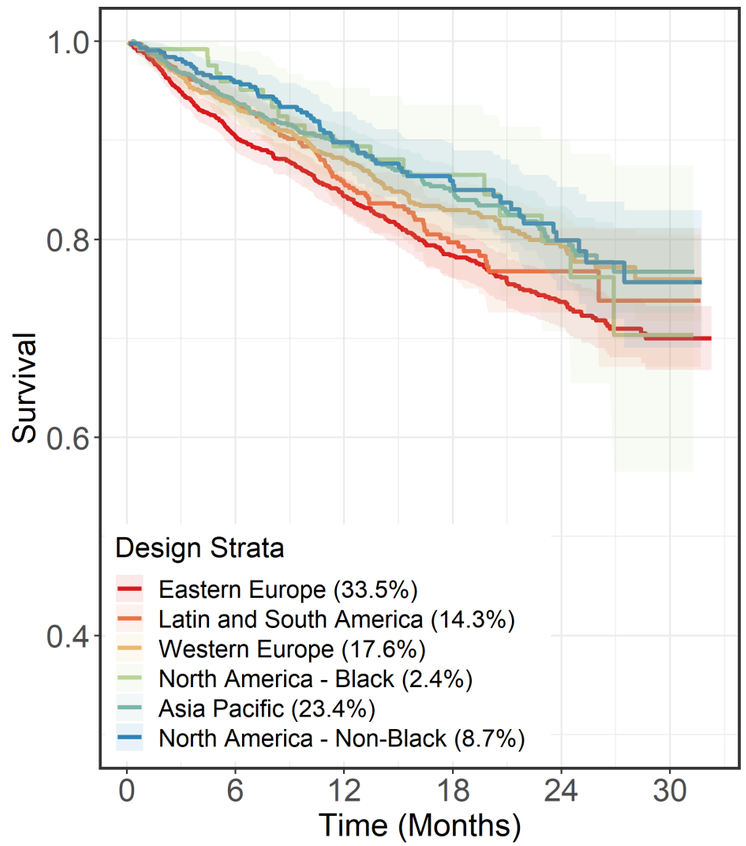
Vericiquat (test treatment) vs. placebo (control treatment)

Kaplan-Meier curves by treatment

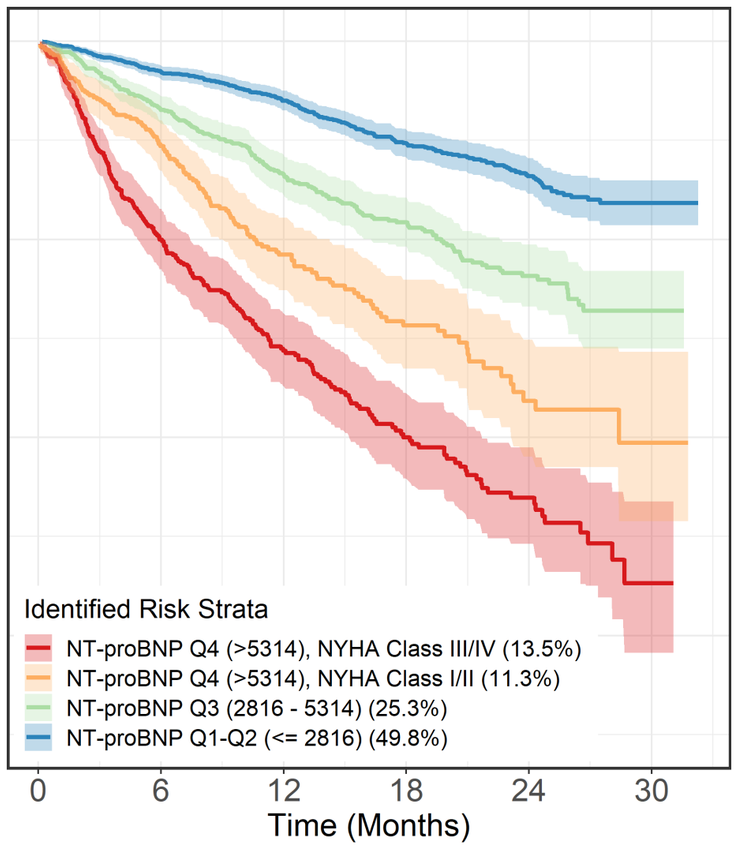


Kaplan-Meier curves by strata

Strata based on pre-specified stratification factors (***design strata***)



Strata identified using treatment-blinded algorithm (***risk strata***)



	HR Estimate (95% CI)	P-value (1-tailed)
Traditional analysis (uses design strata)	0.93 (0.81, 1.06)	0.134
5-STAR analysis (uses risk strata)	0.83 (0.71, 0.97)	0.008

Disclaimer: retrospective analyses for illustration