Duke Clinical Research Institute

Weighted Composite Endpoints Approach: An Alternative to Win Ratio

Huiman Barnhart, PhD
Duke University

Society for Clinical Trials Annual Meeting
May 18, 2026

Win Ratio (WR)

European Heart Journal
doi:10.1093/eurheartj/ehr352

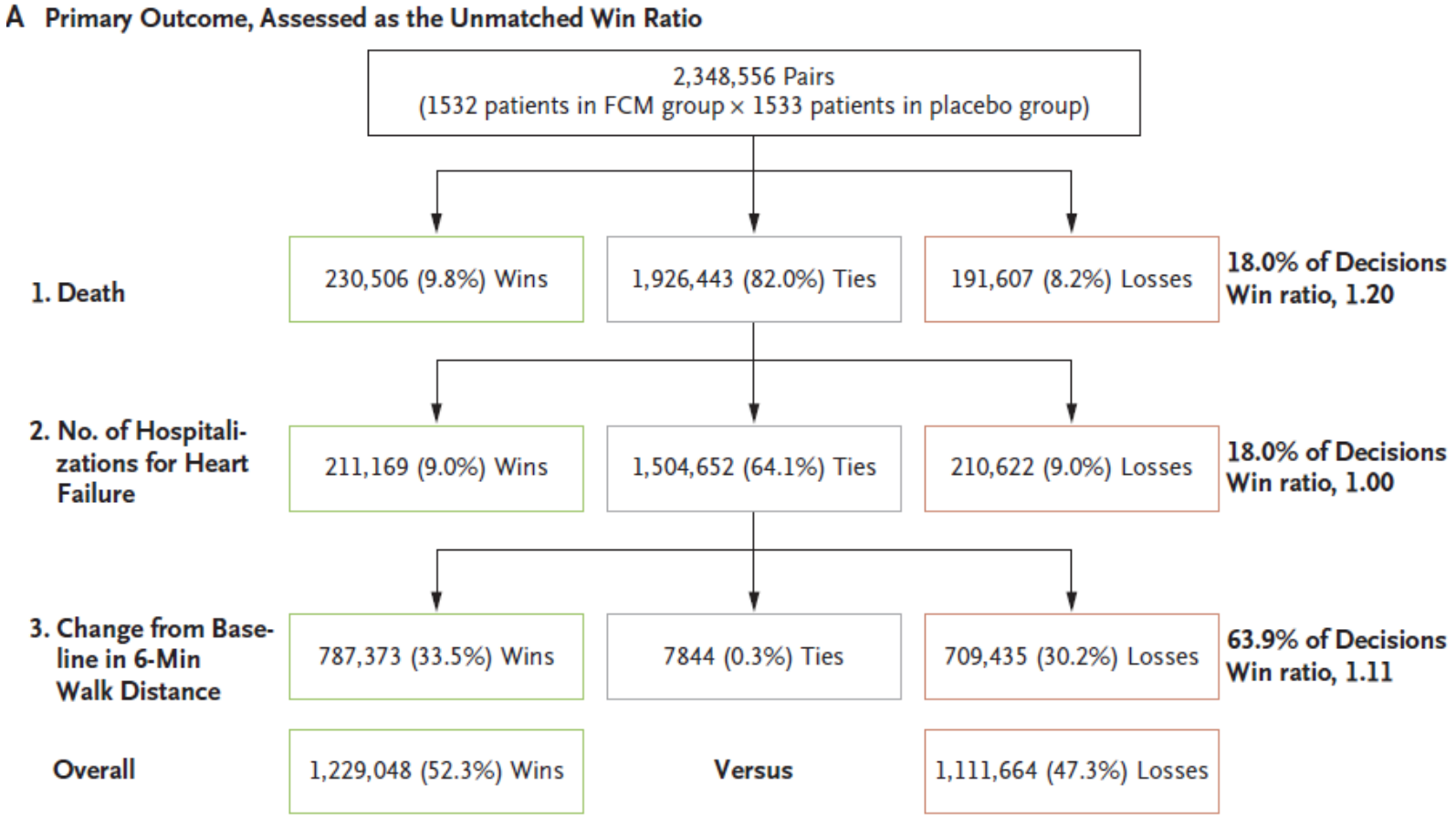
SPECIAL ARTICLE

The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities

Stuart J. Pocock*, Cono A. Ariti, Timothy J. Collier, and Duolao Wang

- Hierarchical endpoints, $Y_1, \ldots, Y_K$, that can be either time to event, counts, binary, or ordinal, are compared for each pair of patients from two treatment groups (A and B) over standard follow-up in a sequence.
- If a patient from group A (B) has a better endpoint than a patient from group B (A), then the pair is a "win" ("loss") for treatment A.
- If a win/loss cannot be determined, patients are compared on the next endpoint in the hierarchy. If win/loss cannot be determined for all endpoints, the pair is a "tie".
- Win Ratio = $\frac{Prob\ (Win)}{Prob(Loss)}$

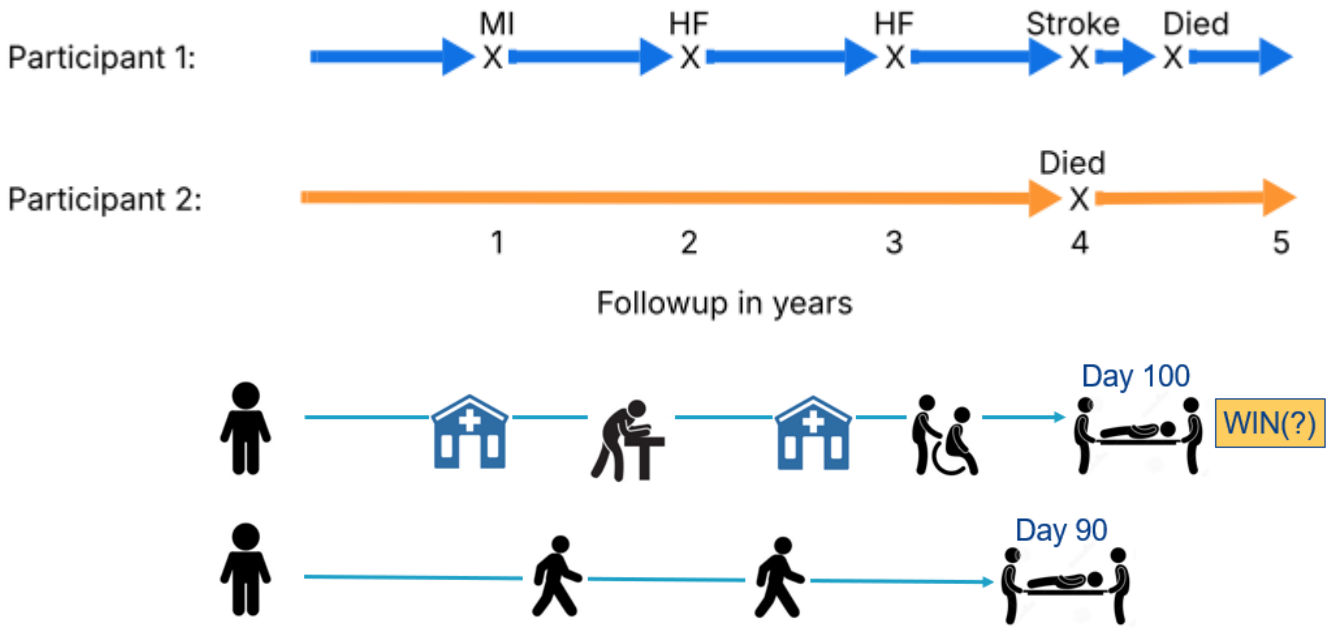
Example of Win Ratio: HEART-FID trial – Majority of wins/loses from 6MWD



Unmatched win ratio (based on the first imputed data set)=(total wins)/(total losses)=1,229,048/1,111,664=1.11 (99% CI, 0.99–1.23)
Overall unmatched win ratio, 1.10 (99% CI, 0.99–1.23; P=0.02)

Why do we need an alternative approach to Win Ratio?

- can handle multiple events/recurrent events
  - But not all events/all patient's experience are utilized, e.g., MI < HF < Stroke < Death
    Participant 1 wins over Participant 2. MI, HFs and Stroke in participant 1 are ignored



Why do we need an alternative approach to Win Ratio?

- In Win Ratio approach, a win/lose is a win/lose. It doesn't matter on which endpoint you win/lose. So, it treats win on first endpoint and subsequent conditional wins equally. One endpoint can still potentially dominate.

$$WR(S; Y_1, \ldots, Y_K) = \frac{P(\text{A wins on } Y_1) + P(\text{A wins on } \text{Y}_2, \text{ties on } \text{Y}_1) + \cdots + P(\text{A wins on } \text{Y}_\text{K}, \text{ties on } \text{Y}_1 \text{ through } \text{Y}_{\text{K}-1})}{P(\text{A loses on } Y_1) + P(\text{A loses on } \text{Y}_2, \text{ties on } \text{Y}_1) + \cdots + P(\text{A loses on } \text{Y}_\text{K}, \text{ties on } \text{Y}_1 \text{ through } \text{Y}_{\text{K}-1})}$$

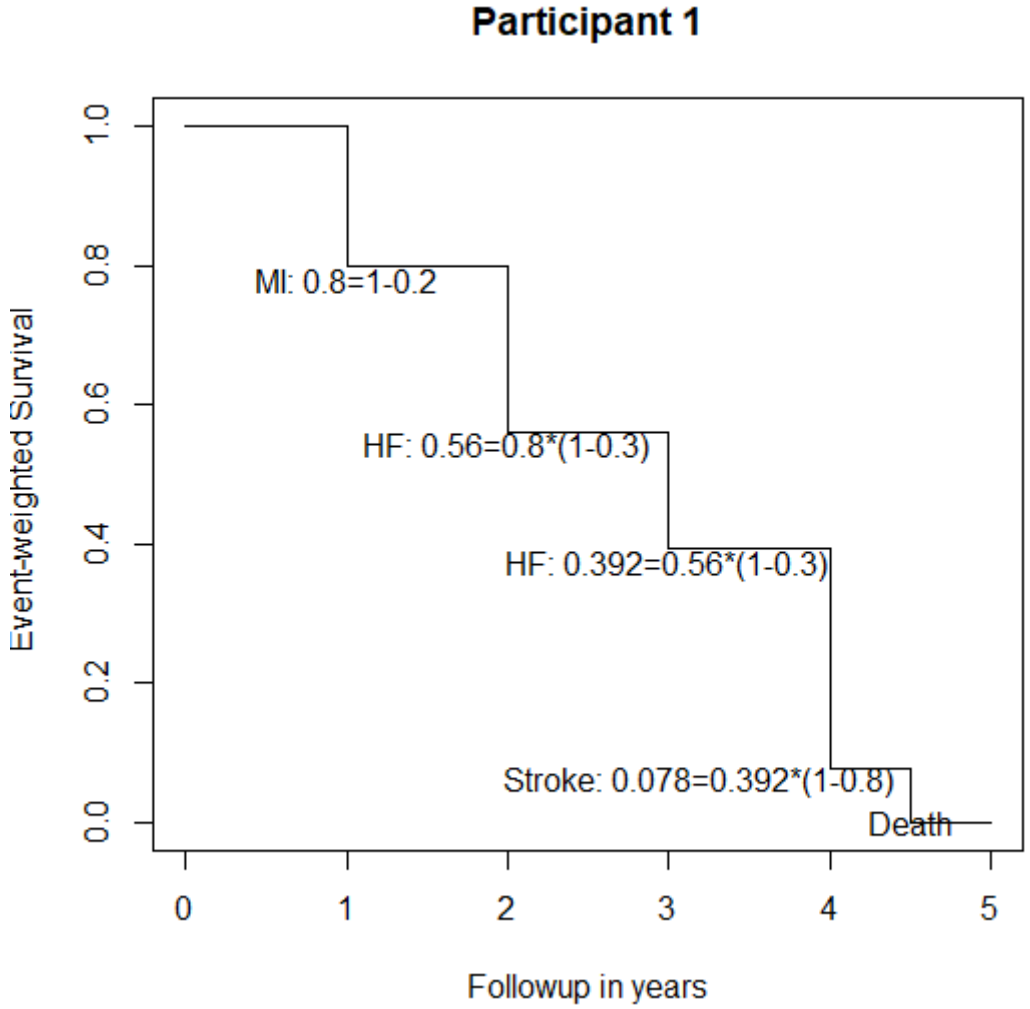
Weighted Composite Endpoints Approach as an Alternative

Bakal et al. *European Heart Journal* 2014; Bakal et al. *SMMR*, 2015; Nabipoor et al. *BMC Medical Research Methodology*, 2023

- Uses all events that occur to a patient during the study period, at the time of occurrence
- Each subject starts with survival of 1.0.
- An occurrence of an event reduces the **weighted** survival **multiplicatively** by an amount based on the relative weight/damage assigned to the type of event

Weights scaled relative to death:

- Death – 1
- Stroke – 0.8
- HF – 0.3
- MI – 0.2

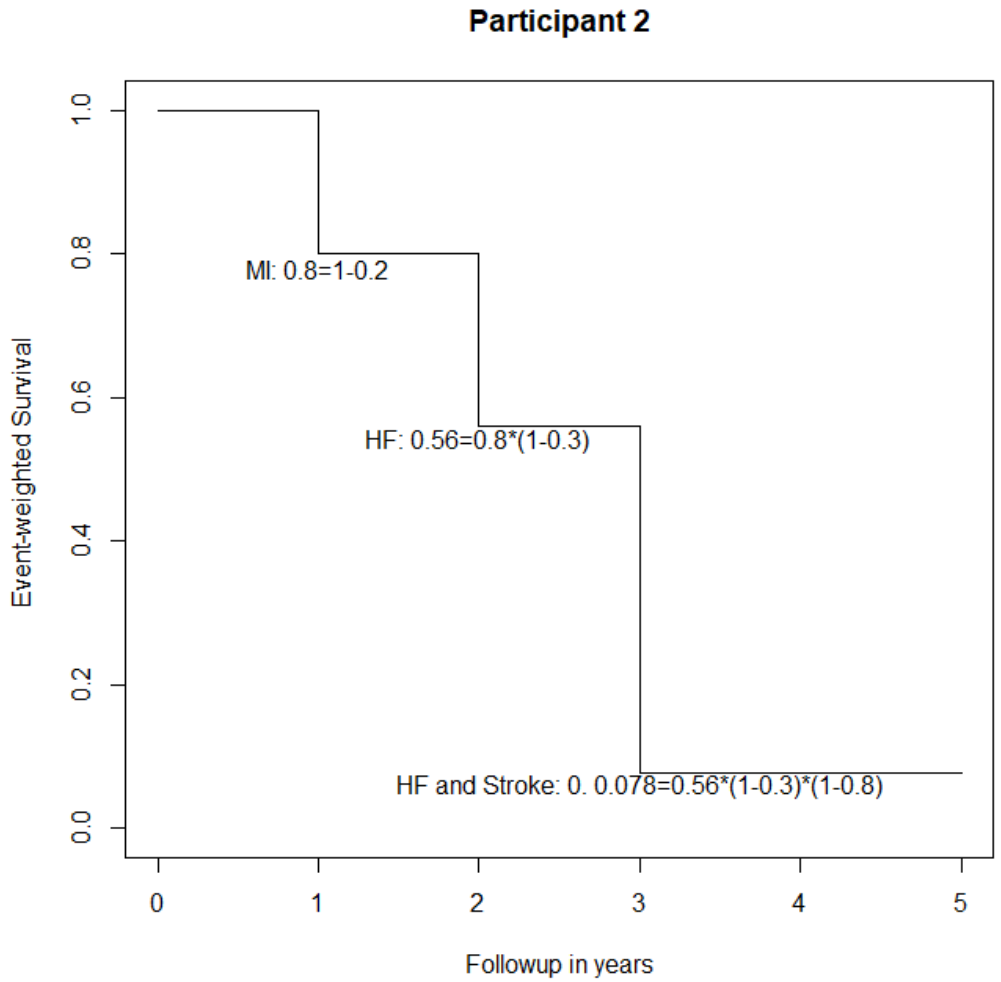


What happens if there are two events occurred at the same day?

- Let's say this participant had HF at year 3=1095 days and then stroke at the same day. The weights for HF and Stroke together is 0.3+0.8 > 1.0.
- If HF occurred first followed by stroke, the event-weighted survival at year 3=0.56*(1-0.3)*(1-0.8)=0.078
- If stroke occurred first followed by HF, the event-weighted survival at year 3=0.56*(1-0.8)*(1-0.3)=0.078

Weights scaled relative to death:

- Death – 1
- Stroke – 0.8
- HF – 0.3
- MI – 0.2



How do you decide on Weights?

- Use a Delphi Panel consisting of clinical experts in the area

- Patient Preferences via Patient preference survey

Weighted-event Survival Probability - Estimation

Bakal et al. European Heart Journal 2014; Bakal et al. SMMR, 2015; Nabipoor et al. BMC Medical Research Methodology, 2023

- Let $t_0 = 0 < t_1 < \ \ldots < t_j < \ \ldots < t_J$ be the finite set of distinct observed event times
- K possible type of events that can occur for each subject at each time t_j
- Weight vector: $W = (w_1, \ldots, w_k, \ldots, w_K)^\top, w_1 = 1 \text{ and } w_k < 1$
- E_{ijk} denotes event indicator for type k at distinct time t_j for subject i

Existing Estimation of Event-weighted Survival without Censoring

If no censoring is present, the estimated weighted survival probability for subject *i*

$$\hat{S}_i(t) = \prod_{j:t_j \le t} (1 - w_{ij})$$

$$\omega_{ij} = \sum_{k=1}^{K} w_k E_{ijk} \quad \text{for subject } i \text{ at time } \mathrm{t_j}$$

- **Estimated Weighted Kaplan-Meier Curve** (note that both d_j and n_j are not necessarily integers)

$$\hat{S}(t_j) = \prod_{m \le j} \left(1 - \frac{d_m}{n_m}\right) = \frac{1}{n} \sum_{i=1}^{n} \hat{S}_i(t)$$

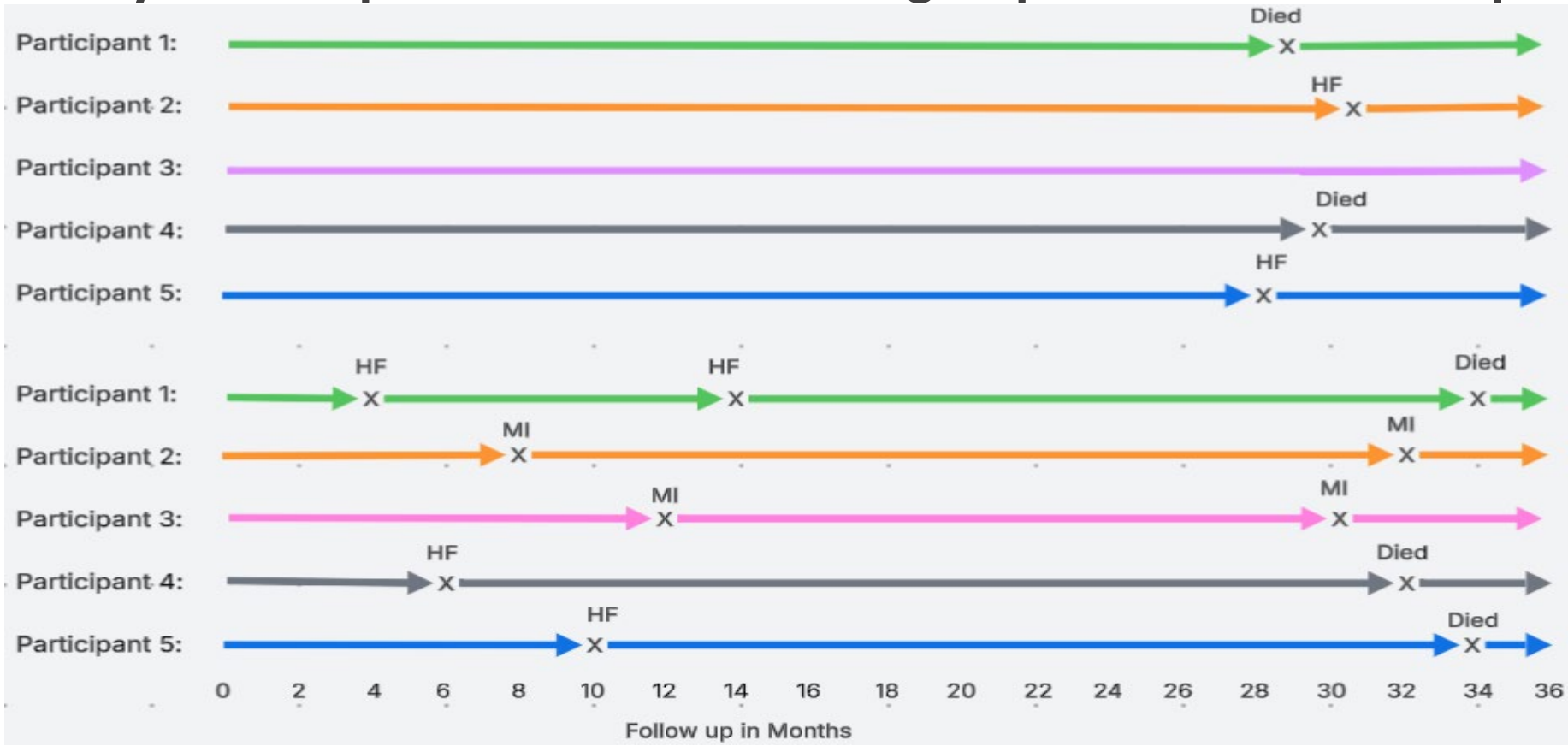
  - Weighted number of death-like events: $d_j = \sum_{i=1}^{n} [\hat{S}_i(t_{j-1}) - \hat{S}_i(t_j)]$
  - Weighted number at risk: $n_j = \sum_{i=1}^{n} \hat{S}_i(t_{j-1})$

Working Example: WR=11/14=0.79

Consider a study with 10 patients with 5 in each group and 36 M follow-up

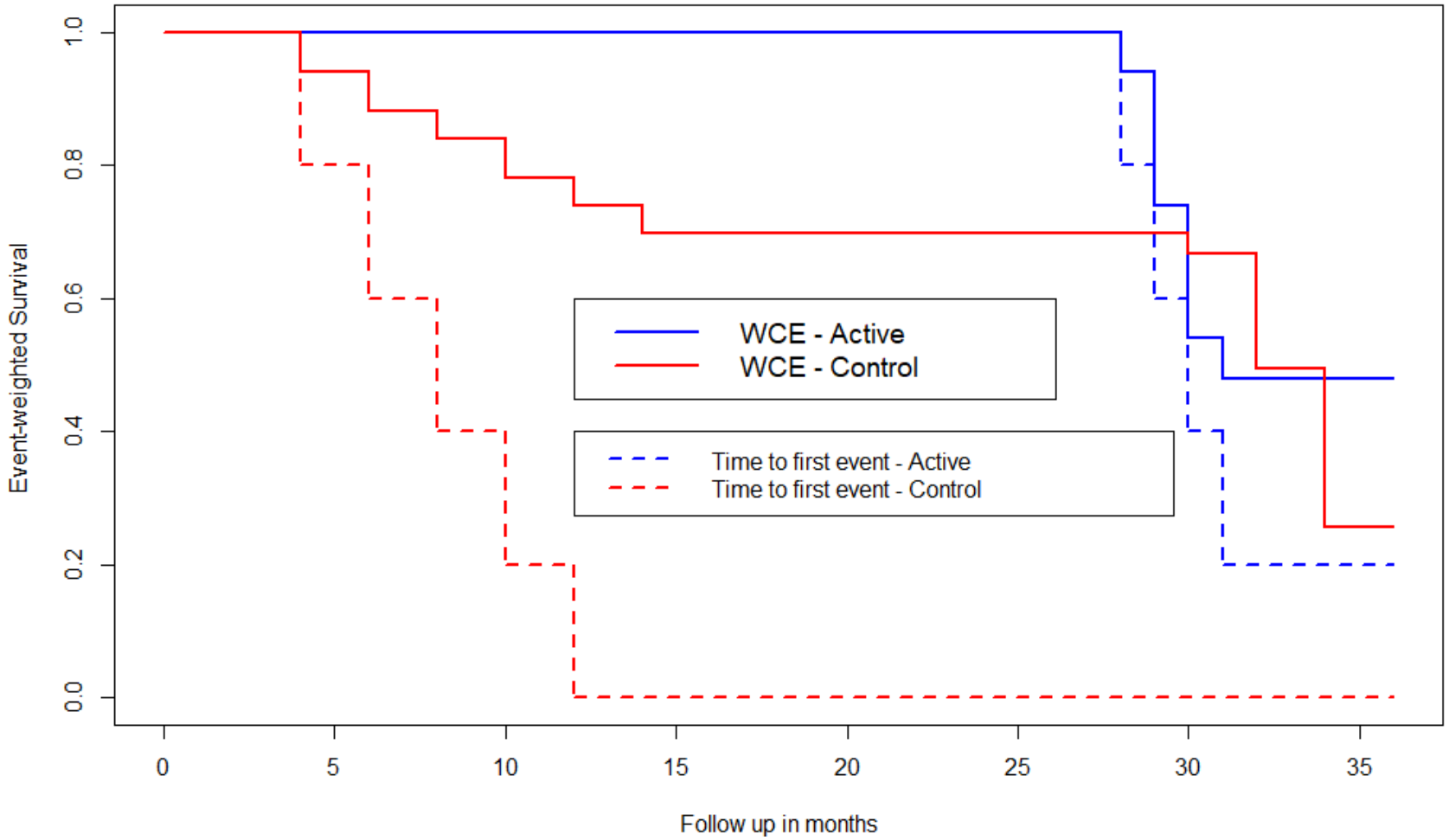
Active Arm

Control Arm



Event Weights: Death=1, HF=0.3, MI=0.2

Working Example 2: WCE vs. Time to First Event vs. WR(=0.79)



Difficulty with the existing WCE approach

- What's the underlying model?
- What is the estimand, the true event-weighted survival?
- Censoring is ubiquitous in clinical trials, how to extend to situation with censoring?
- How to compare the two curves?
- Trial design

Proposed underlying model for Event-Weighted Survival

- Pre-specified time horizon, S.
- A randomly selected subject starts with survival of 1.
- The K types of events are ordered from the most important to the least important with pre-specified weights, $w_1 \geq w_2 \geq \cdots \geq w_k \geq \cdots \geq w_K > 0$ for the K different types of events
- $f_{ak}(t)$ - the probability of k^{th} event occurrence at time *t* in group *a* conditional on survival prior to time t.
  - Note that the weights and probabilities are two separate specifications that do not impact each other, but both impact the event-weighted survival

$$S_a(t) = \prod_{t=0}^{S} (1 - \sum_{k=1}^{K} w_k f_{ak}(t))$$

Proposed Estimation under Censoring

If censoring is present, the estimated weighted survival for subject *i*

$$\hat{S}_i(t_j) = \prod_{j:t_j \le t} (1 - \omega_{ij})$$

Weighted damage: $\omega_{ij} = \begin{cases} \sum_{k=1}^{K} w_k E_{ijk} & \text{if subject } i \text{ is not censored prior to } t_j \\ \sum_{k=1}^{K} \boldsymbol{w_k \hat{p}_{jk}} & \textbf{if subject } \boldsymbol{i} \textbf{ is censored prior to } \boldsymbol{t_j} \end{cases}$

With $\hat{p}_{jk} = \frac{e_{jk}}{n_j^*} = \frac{\#\ observed\ events\ k\ at\ time\ t_j}{\#\ subjects\ alive\ and\ not\ censored\ at\ t_j}$

Estimated Weighted Kaplan-Meier Curve $\hat{S}(t_j) = \prod_{m \le j} \left(1 - \frac{d_m}{n_m}\right) = \frac{1}{n}\sum_{i=1}^{n} \hat{S}_i(t)$

- Weighted number of death-like events: $d_j = \sum_{i=1}^{n} [\hat{S}_i(t_{j-1}) - \hat{S}_i(t_j)]$
- Weighted number at risk: $n_j = \sum_{i=1}^{n} \hat{S}_i(t_{j-1})$

Proposed Variance Estimation under Censoring via Delta Method

$$\mathbf{Var}\left(\widehat{\boldsymbol{S}}\,(\boldsymbol{t})\right) \approx \frac{\mathbf{1}}{\boldsymbol{n^2}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \widehat{\boldsymbol{S}}_{\boldsymbol{i}}^{\mathbf{2}}(\boldsymbol{t}) \sum_{\boldsymbol{j}:\, \boldsymbol{t_j \le t}} \frac{\mathbf{1}}{\left(\mathbf{1} - \boldsymbol{W^T P_j}\right)^2} \boldsymbol{W^T}\left[\mathbf{diag}\left(\boldsymbol{P_j}\right) - \boldsymbol{P_j P_j^T}\right]\boldsymbol{W} \textbf{ + extra term}$$

Where $P_j = \left(p_{j1}, p_{j2}, \ldots, p_{jK}\right)^T$ with $p_{jk} = E(E_{ijk})$

- For construction of 95% confidence bands
- For hypothesis testing on weighted events at a pre-specified time per # patients, e.g., 100 patient.

Example: COMBINE-AF Trials

- A COllaboration between Multiple institutions to Better Investigate Non-vitamin K antagonist oral anticoagulant usE in Atrial Fibrillation
- Consists of five trials:

  RELY, ROCKET-AF, ARISTOTLE, AVERROES and ENGAGE

- **A sub-study with four trials** (exclude AVERROES trial due to comparison to aspirin rather than warfarin)
  - **To compare standard direct oral anticoagulants (DOACs) versus warfarin** via WCE using data from four trials: RE-LY, ROCKET-AF, ARISTOTLE, ENGAGE
  - This investigation is led by Satoshi Shoji, MD PhD, Pishoy Gouda, MD

Example: COMBINE-AF

Trial	Sample Size	Median (Q1, Q3) Follow-up
ARISTOTLE	18,201	21.6 (16.1-28.0) months
ENGAGE	14,071	32.0 (29.5-32.0) months
RE-LY	12,098	24.0 (18.9-29.2) months
ROCKET-AF	14,264	21.9 (15.8-27.8) months
Total	58,634	25.4 (18.3, 31.8) months

Example: COMBINE-AF

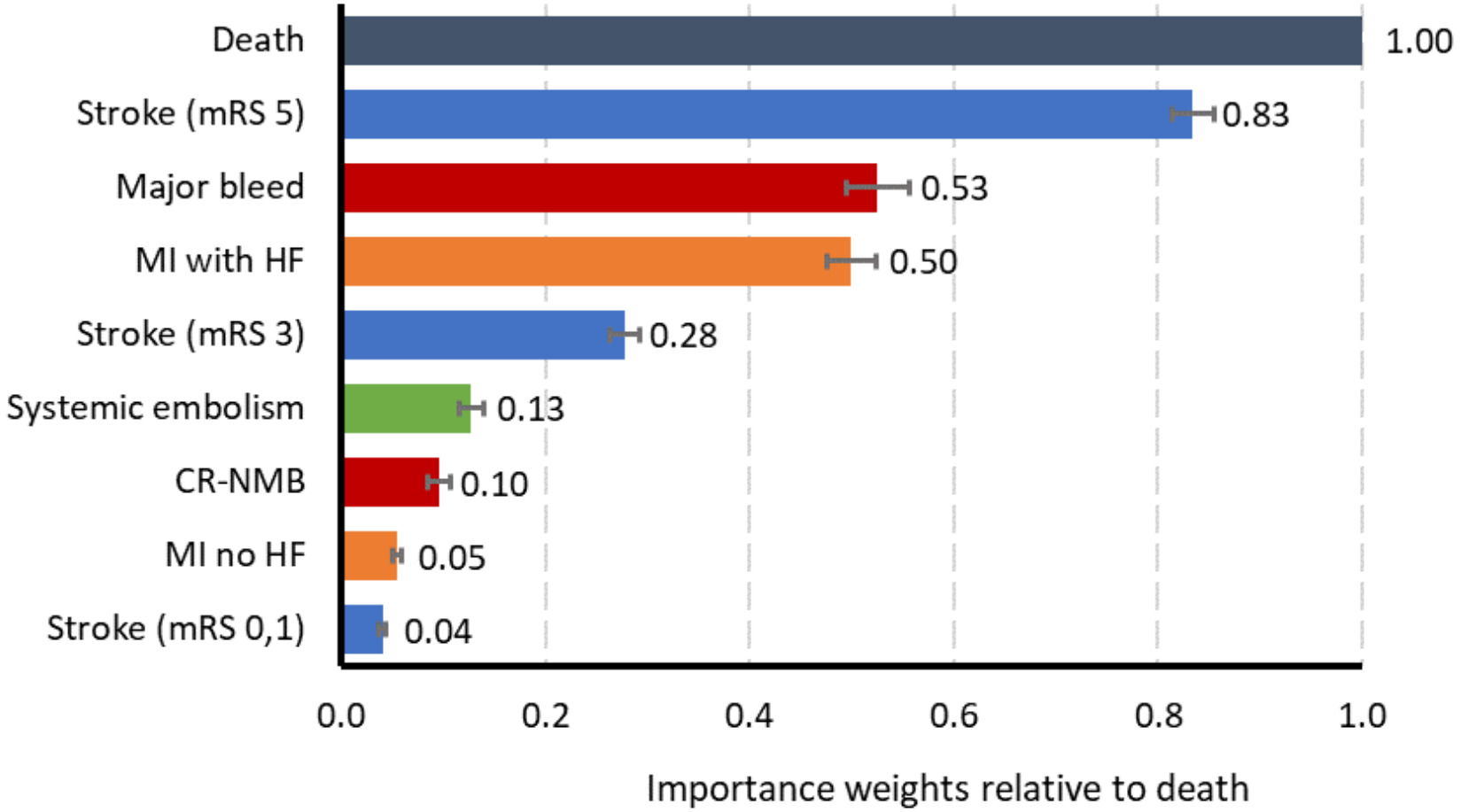
Trial	DOAC	Warfarin
ARISTOTLE	9,120	9,081
ENGAGE AF-TIMI 48	7,035	7,036
RE-LY	6,076	6,022
ROCKET-AF	7,131	7,133
Overall COMBINE AF	29,362	29,272

Example: COMBINE-AF

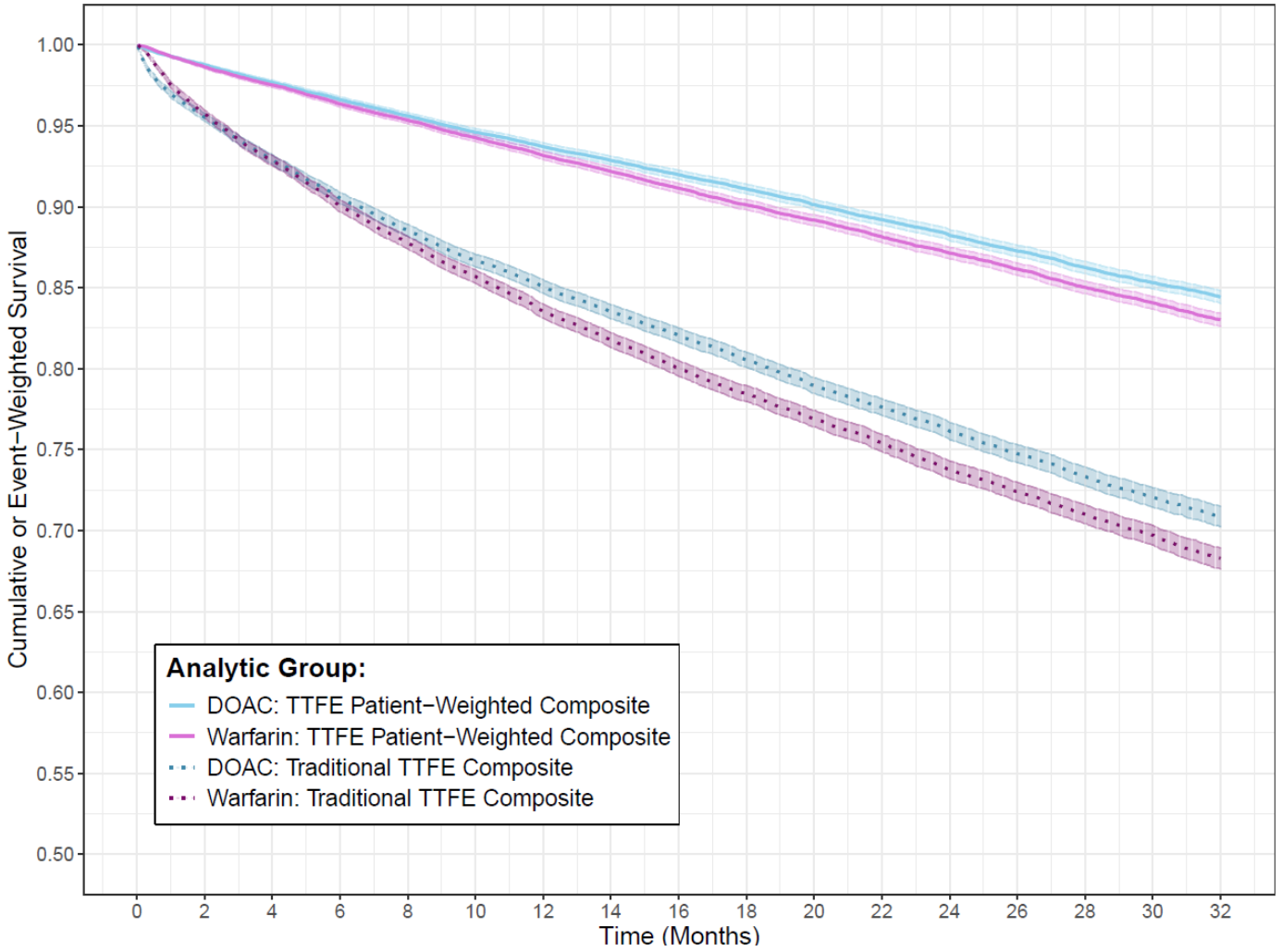
Event	Weight	# Events (N = 58,541)	Avg. Time-To-First Event	Mean (Range) # Events per Patients among those with events
Death from Any Cause	1	4736	429	1 (1,1)
Stroke (mRS 4, 5)	0.83	259	378	1.036 (1,2)
Major Bleeding	0.53	4049	354	1.125 (1,6)
Stroke (mRS 2, 3)	0.28	344	356	1.039 (1, 2)
Systemic Embolism	0.13	152	364	1.041 (1, 2)
Non-Major Clinically Relevant Bleeding	0.10	7681	313	1.352 (1, 13)
Stroke (mRS 0, 1)	0.04	896	370	1.035 (1,2)

Patient-Stated Preference - Relative importance weights

Reed SD, et al. Participant Engagement And pReference study for cLinical outcomes associated with Atrial Fibrillation (PEARL-AF). JACC Adv VOL. 3, NO. 12, 2024



Application to COMBINE-AF: WCE vs. time to first event



Solid lines: event-weighted survival
Dotted lines: time to first event

Death-like events per 100 patients at Year 2
DOACs vs. warfarin: 11.74 vs. 12.85

Difference: –1.11 [95% CI: –1.61 to –0.61]
P < 0.001. Reduction of one death-like event

Application Paper to appear in NEJM Evidence

Discussion: Dependence on Weights

- WCE approach depends on pre-specified weights – a plus on transparency in reporting
  - ➢ All existing methods on composite endpoint or hierarchical endpoints impose weights but not explicitly stated, e.g.,
    - Time to first event use weight of 1 for all events
    - Win ratio puts equal weights on all wins

$$WR(S; Y_1, \dots, Y_K) = \frac{P(\text{A wins on } Y_1) + P(\text{A wins on } \mathrm{Y_2}, \text{ties on } \mathrm{Y_1}) + \cdots + P(\text{A wins on } \mathrm{Y_K}, \text{ties on } \mathrm{Y_1} \text{ through } \mathrm{Y_{K-1}})}{P(\text{A loses on } Y_1) + P(\text{A loses on } \mathrm{Y_2}, \text{ties on } \mathrm{Y_1}) + \cdots + P(\text{A loses on } \mathrm{Y_K}, \text{ties on } \mathrm{Y_1} \text{ through } \mathrm{Y_{K-1}})}$$

$$= w_1\ P(\text{A wins on } Y_1) + w_2 P(\text{A wins on } \mathrm{Y_2}, \text{ties on } \mathrm{Y_1}) + \cdots + w_k P(\text{A wins on } \mathrm{Y_K}, \text{ties on } \mathrm{Y_1} \text{ through } \mathrm{Y_{K-1}})$$

$$w_1 = w_2 = \cdots = w_K = \left(P(\text{A loses on } Y_1) + P(\text{A loses on } \mathrm{Y_2}, \text{ties on } \mathrm{Y_1}) + \cdots + P(\text{A loses on } \mathrm{Y_K}, \text{ties on } \mathrm{Y_1} \text{ through } \mathrm{Y_{K-1}})\right)^{-1}$$

Discussion: Dependence on Weights

- Weights may vary from patient to patient – can I use my own weights?
  - With published work, provide a web link that allow people to put their own weights - to generate results without any identifiable information.
- How do you compare results from studies with different weights?
  - Generate results based on the same weights and same FU time for comparison – may be achieved with above links.
- It may be difficult to get consensus on a unique set of weights
  - Establish consensus weight repositories for different populations and diseases
  - Policy makers need to weigh in both clinical and patients' perspective

Take Home Messages

- WCE produces Kaplan-Meier-like survival curves that are intuitive and easy to visualize
- WCE uses all information longitudinally, unlike Win Ratio that only use the worst outcome in paired comparisons
- WCE provides absolute measure rather than relative measure and thus it's easier to interpret
- WCE depends on pre-specified weights for setting hierarchy with relative magnitudes
  - Not all wins are equal, and it provides absolute trade off between different events
- Weights can be determined based on patients' preference – a plus for patient centered approach

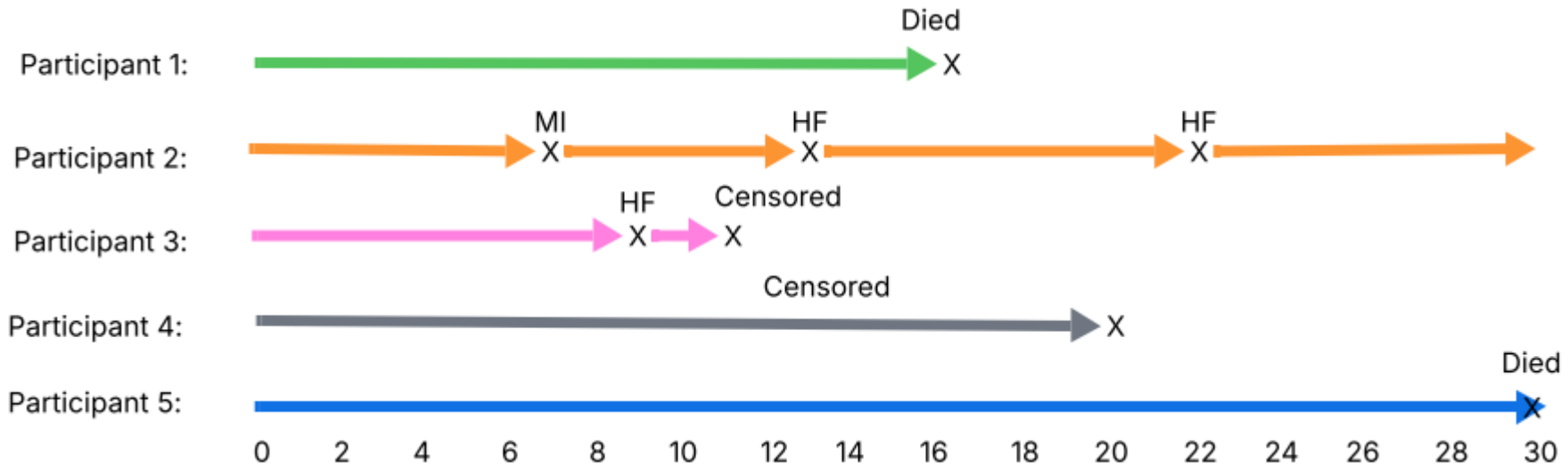
Future Work

- Comparison of two weighted survival curves – to be developed
- Extension to non-time-to-event endpoints – to be developed

Questions and Comments

Working Example

Consider a study with 5 patients and a total 30-day follow-up period



Distinct Event Times: 7, 9, 13, 16, 22, 30

Event Weights: Death=1, HF=0.3, MI=0.2

Working Example 1

	Event Timepoints						
ID	0	7	9	13	16	22	30
1	1.00	1.00	1.00	1.00	0.00	0.00	0.00
2	1.00	0.80	0.800	0.560	0.560	0.392	0.392
3	1.00	1.00	0.700	**0.6475**	**0.4856**	**0.4128**	**0.2064**
4	1.00	1.00	1.00	1.00	1.00	**0.850**	**0.425**
5	1.00	1.00	1.00	1.00	1.00	1.00	0.00
Average	**1.00**	**0.96**	**0.90**	**0.842**	**0.609**	**0.531**	**0.205**

For subject 3 censored at day 12
At Day 13: 0.6475=0.7*(1-0.3*(1/4)) - one HF over 4 subjects
At Day 16: 0.4856=0.6475*(1-1*1/4)) - one death over 4 subjects

At Day 22: 0.4128=0.4856*(1-0.3*(1/2)) – one HF over 2 subjects
At Day 30: 0.1785=0.4128*(1-1*1/2) – one death over 2 subjects

`WRNet`: Regularized win ratio regression

Variable selection and risk prediction with hierarchical composite outcomes

Lu Mao

Department of Biostatistics & Medical Informatics

University of Wisconsin-Madison

Introduction

Hierarchical Composite Endpoints

- **Composite outcomes**
  - **Components**: Death, hospitalization, other events
  - **Standard approach**: Time to first event (e.g., Cox model)
- **Win ratio (WR)** (Pocock et al., 2012)
  - **Pairwise comparisons**: Treatment vs control (Buyse, 2010)
  - **Each pair**: Death > hospitalization (> other events)
  - **Effect size**

$$\text{WR} = \frac{\text{Number of wins}}{\text{Number of losses}}$$

Data and Notation

- **Outcomes data** (subject i)
  - D_i: time to death
  - T_i: time to first nonfatal event
  - $\boldsymbol{Y}_i(t) = (D_i \wedge t, T_i \wedge t)$: cumulative data at t
    - $x \wedge y = \min(x, y)$
- **Win indicator**

$$\mathcal{W}(\boldsymbol{Y}_i, \boldsymbol{Y}_j)(t) = \underbrace{I(D_j < D_i \wedge t)}_{\text{Win on survival}} + \underbrace{I(D_i \wedge D_j > t,\ T_j < T_i \wedge t)}_{\text{Tie on survival, win on nonfatal ev}}$$

Win Ratio Regression

- **Semiparametric regression**

$$\frac{\mathrm{pr}\{\mathcal{W}(\boldsymbol{Y}_i, \boldsymbol{Y}_j)(t) = 1 \mid z_i, z_j\}}{\mathrm{pr}\{\mathcal{W}(\boldsymbol{Y}_j, \boldsymbol{Y}_i)(t) = 1 \mid z_j, z_i\}} = \exp\{\beta^{\mathrm{T}}(z_i - z_j)\}$$

  - z: p-dimensional covariates
  - **Proportional win-fractions (PW) model**
    - Equivalent to Cox PH model in univariate case (Oakes, 2016)—WR = 1/HR
  - $\exp(\beta)$: win ratios associated with unit increases in z

- **Limitation**: $p << n$

Goals

- **Objectives**:
  - Automate variable selection in WR regression
  - Enhance prediction accuracy and model parsimony
- **Approach**:
  - Apply **elastic net** regularization to PW model
    - Mixture of L_1 (lasso) and L_2 (ridge) penalties
  - Implemented in `wrnet` R package

Methods

Standard Model-Fitting

- **Observed data**
  - $\boldsymbol{Y}_i(X_i)$: outcomes data up to $X_i = D_i \wedge C_i$ (C_i: censoring time)
- **Estimating equation** (Mao & Wang, 2021)

$$U_n(\beta) = |\mathcal{R}|^{-1} \sum_{(i,j)\in\mathcal{R}} z_{ij} \left\{ \delta_{ij} - \frac{\exp(\beta^\top z_{ij})}{1+\exp(\beta^\top z_{ij})} \right\}$$

  - $z_{ij} = z_i - z_j$: covariate difference
  - $\delta_{ij} = \mathcal{W}(\boldsymbol{Y}_i, \boldsymbol{Y}_j)(X_i \wedge X_j)$: observed win indicator
  - $\mathcal{R} = \{(i,j) : \delta_{ij} + \delta_{ji} > 0\}$: set of *comparable* pairs

Estimation

- **Estimator**: solving

$$U_n(\hat{\beta}) \quad = 0$$

  - Standard Newton–Raphson algorithm

- **Connection with logistic regression**
  - $U_n(\hat{\beta})$ equivalent to logistic score function (no intercept)
  - Each comparable (i, j) pair as an *observation*
    - δ_{ij}: binary response
    - z_{ij}: covariates

Regularized PW model

- **Objective function** (Zou & Hastie, 2005):

$$l_n(\beta;\lambda) = -|\mathcal{R}|^{-1} \sum_{(i,j)\in\mathcal{R}} [\delta_{ij}\beta^{\mathrm{T}} z_{ij} - \log\{1+\exp(\beta^{\mathrm{T}} z_{ij})\}] + \lambda\left\{(1-\alpha)||\beta||_2^2/2 + \alpha||\beta||_1\right\}$$

  - **Pathwise solution** $\hat{\beta}(\lambda) = \arg\min_\beta l_n(\beta;\lambda)$
    - Numerically equivalent to regularized logistic regression
  - Tuning parameter $\lambda \geq 0$—determined by cross-validation (CV)
    - $\partial l_n(\beta;0)/\partial\beta = U_n(\beta)$
  - Mixing parameter $\alpha \in (0,1)$
    - $\alpha > 0 \longrightarrow$ some components of $\hat{\beta}(\lambda) = 0$ (performs variable selection)

Pathwise Solution

- **Pathwise algorithm** (Friedman et al., 2010)
  - Efficient computation of $\hat{\beta}(\lambda)$ for all λ

```
glmnet::glmnet(x, y, family = "binomial", intercept = FALSE, lambda)
```

  - `x`: covariate matrix containing z_{ij} ,
  - `y`: response vector δ_{ij}
  - `intercept = FALSE` removes intercept
  - `lambda`: user-specified λ vector

Cross Validation

- **CV routine for logistic regression** `cv.glmnet()`
  - Partition *pairs* into k folds—train and validate
  - Built-in `cv.glmnet()`
  - Not appropriate
    - Overlap between analysis and validation sets
    - Inflation of sample size
- **Subject-based CV**
  - Partition subjects into k folds $\mathcal{S}^{(k)}$
  - Train on $\mathcal{S}^{(-k)}$: $\widehat{\beta}^{(-k)}(\lambda) \longrightarrow$ validate on $\mathcal{S}^{(k)}$
  - Identify optimal λ maximizing average concordance index

Win/Risk Score

- **Motivation**
  - Model-predicted win probability given comparability

$$\mu(z_i, z_j; \beta) = \frac{\exp\{\beta^{\mathrm{T}}(z_i - z_j)\}}{1 + \exp\{\beta^{\mathrm{T}}(z_i - z_j)\}}$$

  - $\beta^{\mathrm{T}} z$ measures tendency to win

$$\begin{aligned}
\mu(z_i, z_j; \beta) > 0.5 &\Leftrightarrow \beta^{\mathrm{T}} z_i > \beta^{\mathrm{T}} z_j; \\
\mu(z_i, z_j; \beta) = 0.5 &\Leftrightarrow \beta^{\mathrm{T}} z_i = \beta^{\mathrm{T}} z_j; \\
\mu(z_i, z_j; \beta) < 0.5 &\Leftrightarrow \beta^{\mathrm{T}} z_i < \beta^{\mathrm{T}} z_j.
\end{aligned}$$

  - $-\beta^{\mathrm{T}} z$: risk score

Generalized Concordance Index

- **Validation/test set $\mathcal{S}^*$**
  - Pairwise indices

$$\mathcal{R}^* = \{(i,j) : \delta_{ij} + \delta_{ji} \neq 0; i < j; i,j \in \mathcal{S}^*\}$$

- **Concordance** (Cheung et al., 2019; Harrell et al., 1982; Uno et al., 2011)
  - Proportion of correct ranking of pairs

$$\mathcal{C}(\mathcal{S}^*;\beta) = |\mathcal{R}^*|^{-1} \sum_{(i,j)\in\mathcal{R}^*} \Big[\underbrace{I\{(2\delta_{ij}-1)(\beta^{\mathrm{T}} z_i - \beta^{\mathrm{T}} z_j) > 0\}}_{\text{Concordant pair}} + 2^{-1}\underbrace{I(\beta^{\mathrm{T}} z_i = \beta^{\mathrm{T}} z}_{\text{Tied score}}$$

Validation and Testing

- **Model tuning**
  - kth-fold CV concordance: $C^{(k)}(\lambda) = \mathcal{C}\left(\mathcal{S}^{(k)}; \widehat{\beta}^{(-k)}(\lambda)\right)$
  - **Optimal** $\lambda_{\text{opt}} = \arg\max_{\lambda} K^{-1} \sum_{k=1}^{K} \mathcal{C}^{(k)}(\lambda)$
  - **Final model:** $\widehat{\beta}(\lambda_{\text{opt}})$
- **Variable importance**
  - Component-wise $|\beta| / \text{sd}(z)$
- **Test C-index**
  - $\mathcal{C}(\mathcal{S}^*; \widehat{\beta}(\lambda_{\text{opt}}))$ for test set $\mathcal{S}^*$

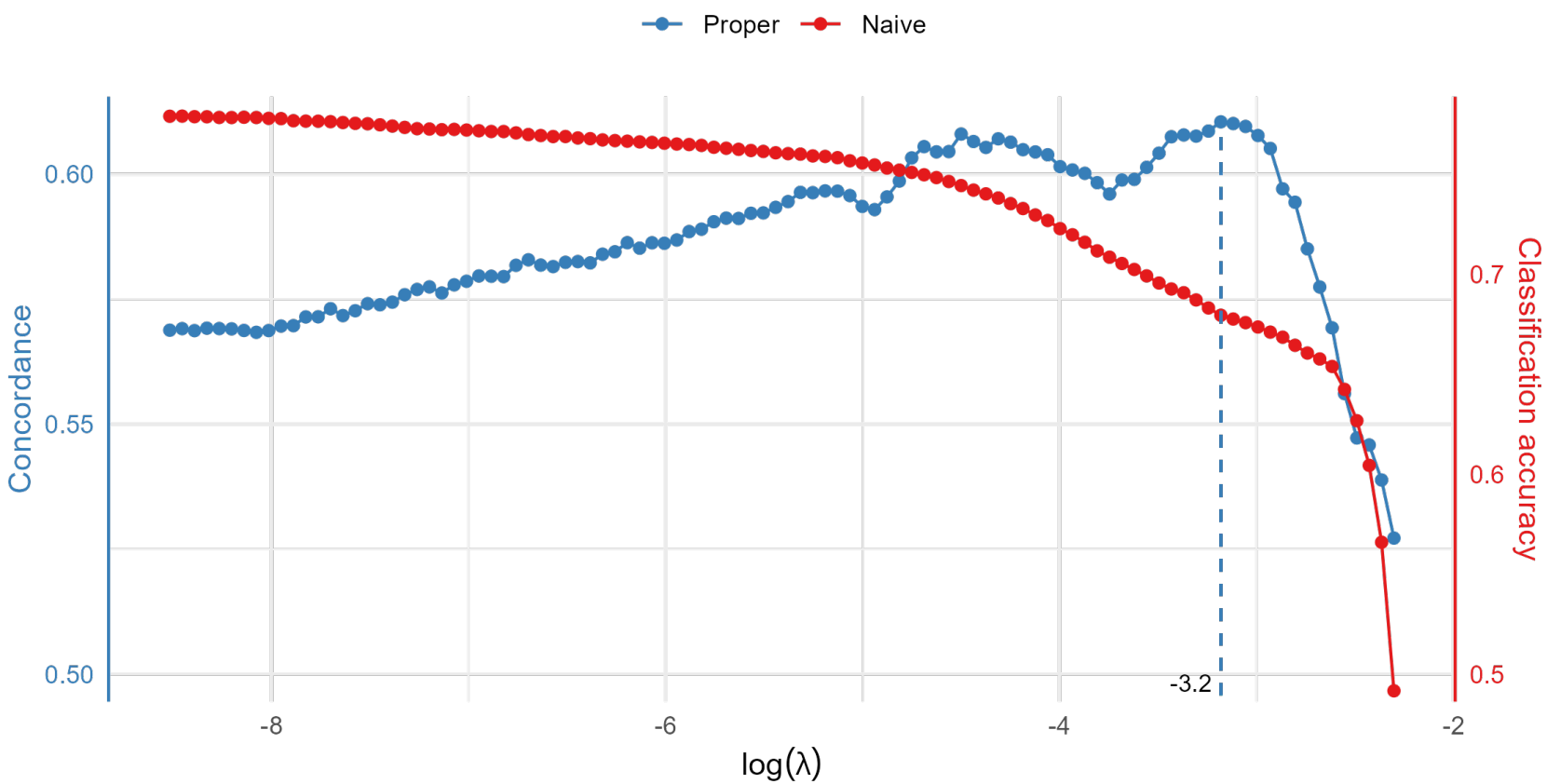
HF-ACTION Application

Study Information

- **HF-ACTION** (O'Connor et al., 2009)
  - **Population**: 2,331 HFrEF patients across North America and France
  - **Objective**: Evaluate the effect of exercise training on composite of death and hospitalization
- **Subgroup of high-risk patients**
  - $n = 426$ high-risk patients (CPX duration ≤ 9 min)
  - Outcomes: death > hospitalization
  - $p = 153$ baseline features
  - Train-test split: 80%/20%

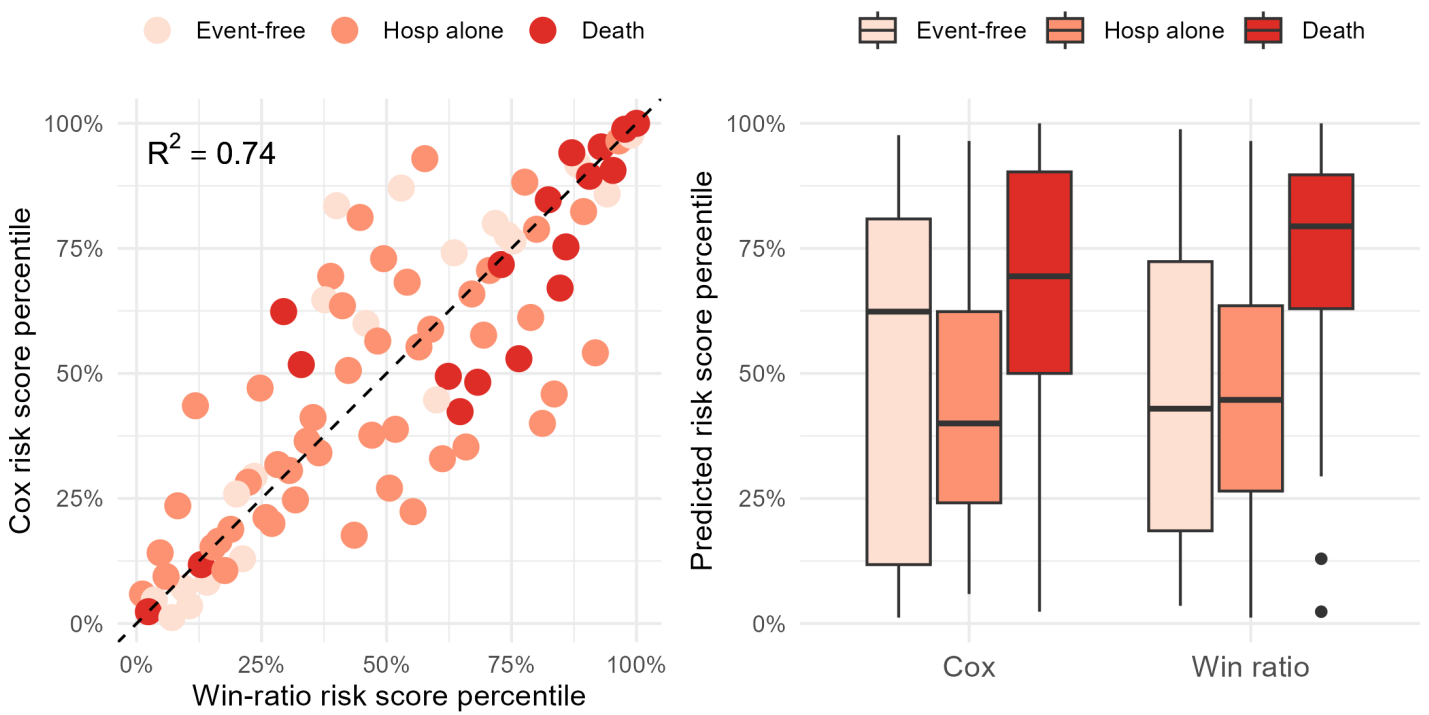
Cross-Validation

- Naive pairwise logistic CV on pairs leads to overfitting



Test Performance - Risk Scores

- **WRNet vs regularized Cox**
  - WRNet better stratifies high-risk patients, especially on mortality



Test Performance - C-Index

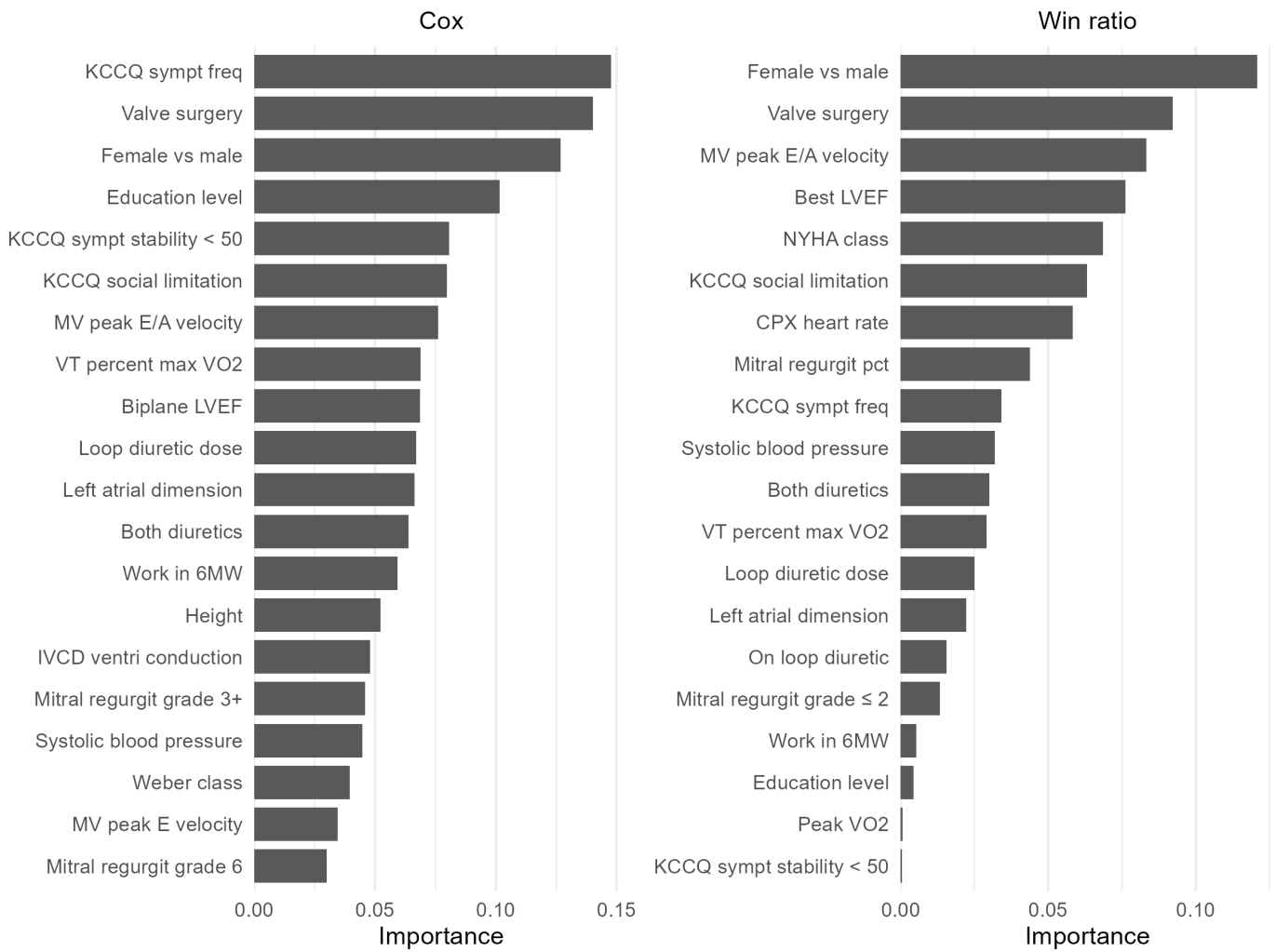
- **WRNet vs regularized Cox**
  - WRNet outperforms Cox on overall and event-specific C-indices

Table 2 Test C-indices of regularized win ratio and Cox models ($\alpha = 1$).

	Cox	Win ratio
Death	0.680	0.735
Hosp	0.522	0.543
Overall	0.572	0.605

Variable Importance

- WRNet selects 20 interpretable variables



Final WR Model

Table 3 Final win ratio model refitted on entire data without regularization.

	Win ratio	95% CI	p-value
Demographics			
Female vs male	1.89	[1.42, 2.53]	<.001
Education level (1-6)	1.07	[0.98, 1.17]	0.153
Medical History			
Valve surgery (y vs n)	0.49	[0.29, 0.81]	0.005
On loop diuretic (y vs n)	0.89	[0.61, 1.30]	0.547
Loop diuretic dose (100mg)	0.91	[0.81, 1.01]	0.080
On both diuretics (y vs n)	0.69	[0.44, 1.10]	0.119
Functional Measurements			
KCCQ symptoms freq	1.01	[1.00, 1.01]	0.104
KCCQ social limitation score	1.00	[1.00, 1.01]	0.349
KCCQ symptom stability < 50 (y vs n)	0.75	[0.51, 1.11]	0.152
Work in 6MW (100kJ)	1.14	[1.00, 1.30]	0.045
NYHA class (III/IV vs II)	0.99	[0.74, 1.33]	0.964
Systolic blood pressure (mmHg)	1.01	[1.00, 1.01]	0.155
CPX Parameters			
VT percent max VO2 (%)	0.15	[0.03, 0.81]	0.028
CPX heart rate (bpm)	1.00	[0.99, 1.01]	0.540
Peak VO2 (mL/kg/min)	1.03	[0.98, 1.08]	0.269
Echocardiographic or MR Measurements			
MV peak E/A velocity (m/s)	0.85	[0.72, 1.01]	0.057
Best LVEF (%)	1.02	[1.00, 1.04]	0.066
Left atrial dimension (mm)	0.91	[0.77, 1.09]	0.318
Mitral regurgitation percent (%)	0.99	[0.98, 1.00]	0.019
Mitral regurgitation grade 2+ (y vs n)	1.35	[0.95, 1.91]	0.095

Summary and Discussion

Summary

- **Methodology**
  - Developed elastic net-regularized WR regression
  - Better aligns with clinical priorities than Cox
  - Accurate feature selection and prediction
- **Tools**
  - `wrnet`: https://lmaowisc.github.io/wrnet/
  - Efficient implementation with `glmnet()` backend
  - Supports CV, test evaluation, and variable importance

Future Topics

- Explore $\alpha \in (0, 1)$ settings
- Extend to time-varying effects and nonlinearities
- Decision trees

References

Buyse, M. (2010). Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. *Statistics in Medicine*, *29*, 3245–3257.

Cheung, L. C., Pan, Q., Hyun, N., & Katki, H. A. (2019). Prioritized concordance index for hierarchical survival outcomes. *Statistics in Medicine*, *38*(15), 2868–2882.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. *JAMA*, *247*(18), 2543–2546.

Mao, L., & Wang, T. (2021). A class of proportional win-fractions regression models for composite outcomes. *Biometrics*, *77*(4), 1265–1275.

O'Connor, C. M., Whellan, D. J., Lee, K. L., Keteyian, S. J., Cooper, L. S., Ellis, S. J., Leifer, E. S., Kraus, W. E., Kitzman, D. W., Blumenthal, J. A., et al. (2009). Efficacy and safety of exercise training in patients with chronic heart failure: HF-ACTION randomized controlled trial. *Journal of the American Medical Association*, *301*(14), 1439–1450.

Oakes, D. (2016). On the win-ratio statistic in clinical trials with multiple types of event. *Biometrika*, *103*(1), 742–745.

Pocock, S., Ariti, C., Collier, T., & Wang, D. (2012). The win ratio: A new approach to the

analysis of composite endpoints in clinical trials based on clinical priorities. *European Heart Journal*, *33*(2), 176–112.

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., & Wei, L.-J. (2011). On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, *30*(10), 1105–1117.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *67*(2), 301–320.

Extending the Win Ratio to Time Spent in a Better Clinical State

Eric Leifer

Office of Biostatistics Research

National Heart, Lung, and Blood Institute

May 18, 2026

Disclaimer

- The views expressed in this talk are mine and do not necessarily represent the views of the National Institutes of Health or the Department of Health and Human Services.

Talk Outline

- Win Ratio background and motivation
- Extensions of the Win Ratio to time spent in a better clinical state
- Discussion

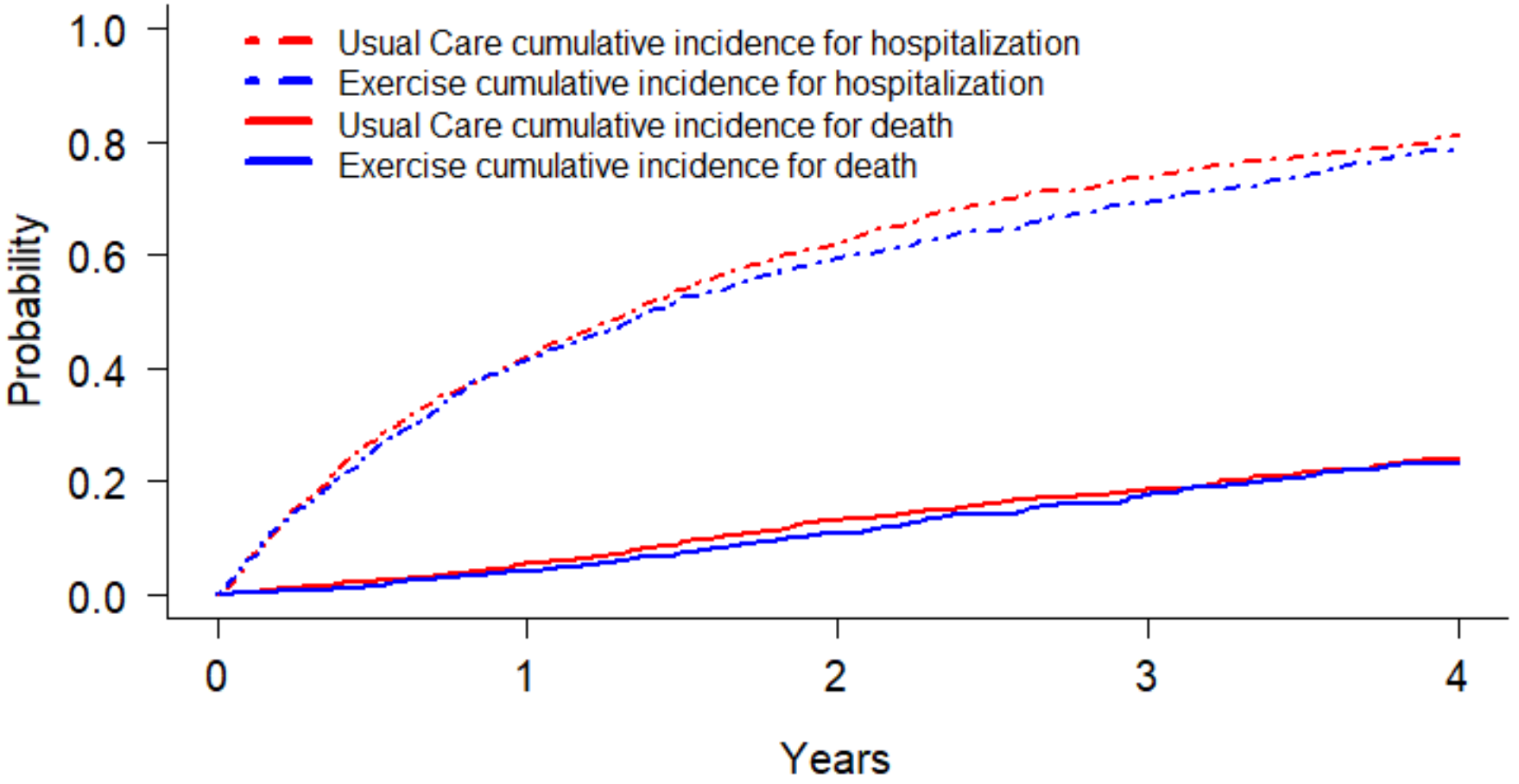
Composite endpoints in heart failure trials

- Heart Failure: A Controlled Trial Investigating Outcomes of Exercise Training (HF-ACTION): 2003-2007
- Randomized over 2000 patients with heart failure to either optimal medical therapy or optimal medical therapy + exercise training
  - Primary endpoint: time-to-death or hospitalization

A key concern with using a time-to-event composite endpoint

- We don’t want treatment benefit on a less serious outcome (hospitalization) to mask treatment harm on a more serious outcome (death)
- That did not happen in HF-ACTION: the treatment and control mortality curves were nearly equivalent.
- The Cox model hazard ratio was in the direction of treatment benefit; 0.93, 95% CI (0.83, 1.03), p = 0.14
  - Due to slightly more favorable hospitalization outcome.
- The win ratio which put more emphasis on the mortality outcome had a less significant p-value (p=0.25).

HF-ACTION Cumulative Incidence Curves for Death, respectively, Hospitalization

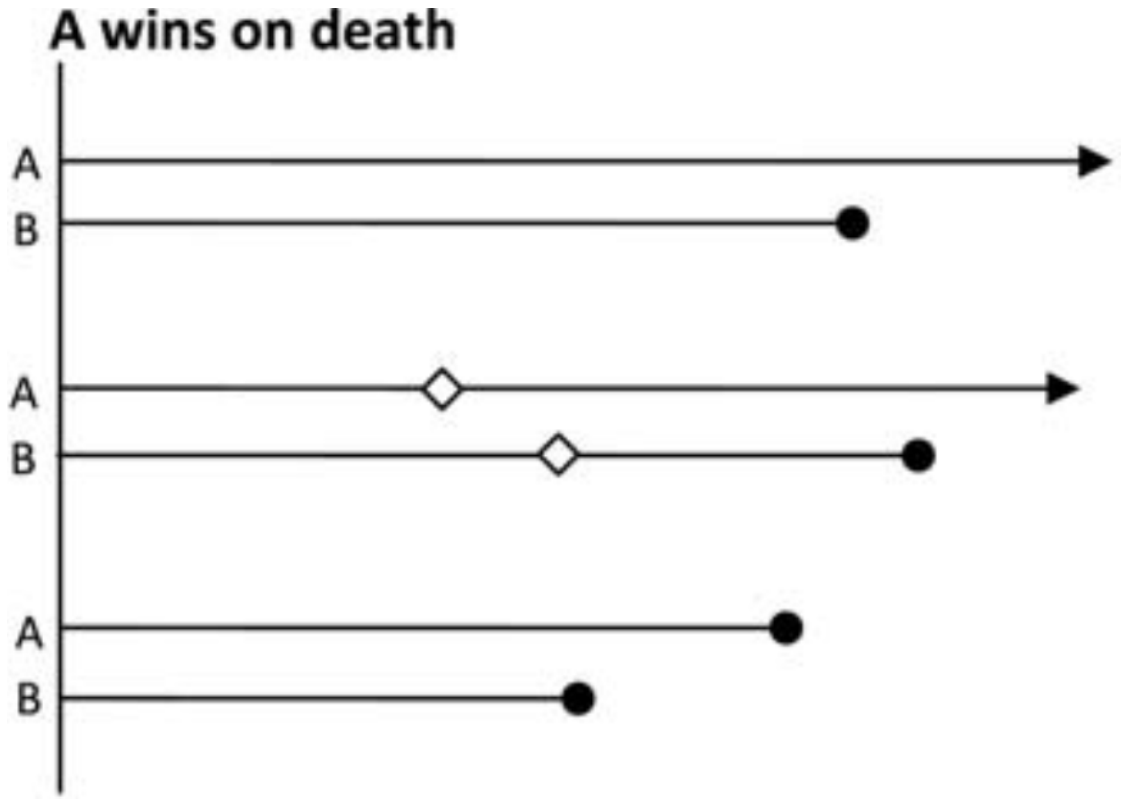


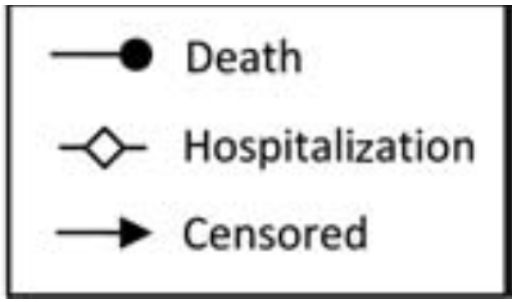
Win Ratio

- **Research Question**: Is treatment A (exercise training) better than treatment B (usual care) with respect to the hierarchical ordering of endpoints?
- **Analysis method**: look at all pairs of A-B subjects and count the number of *A* wins and *B* wins
- **Win Ratio**

$$\frac{number\ A\ wins}{number\ of\ B\ wins}$$

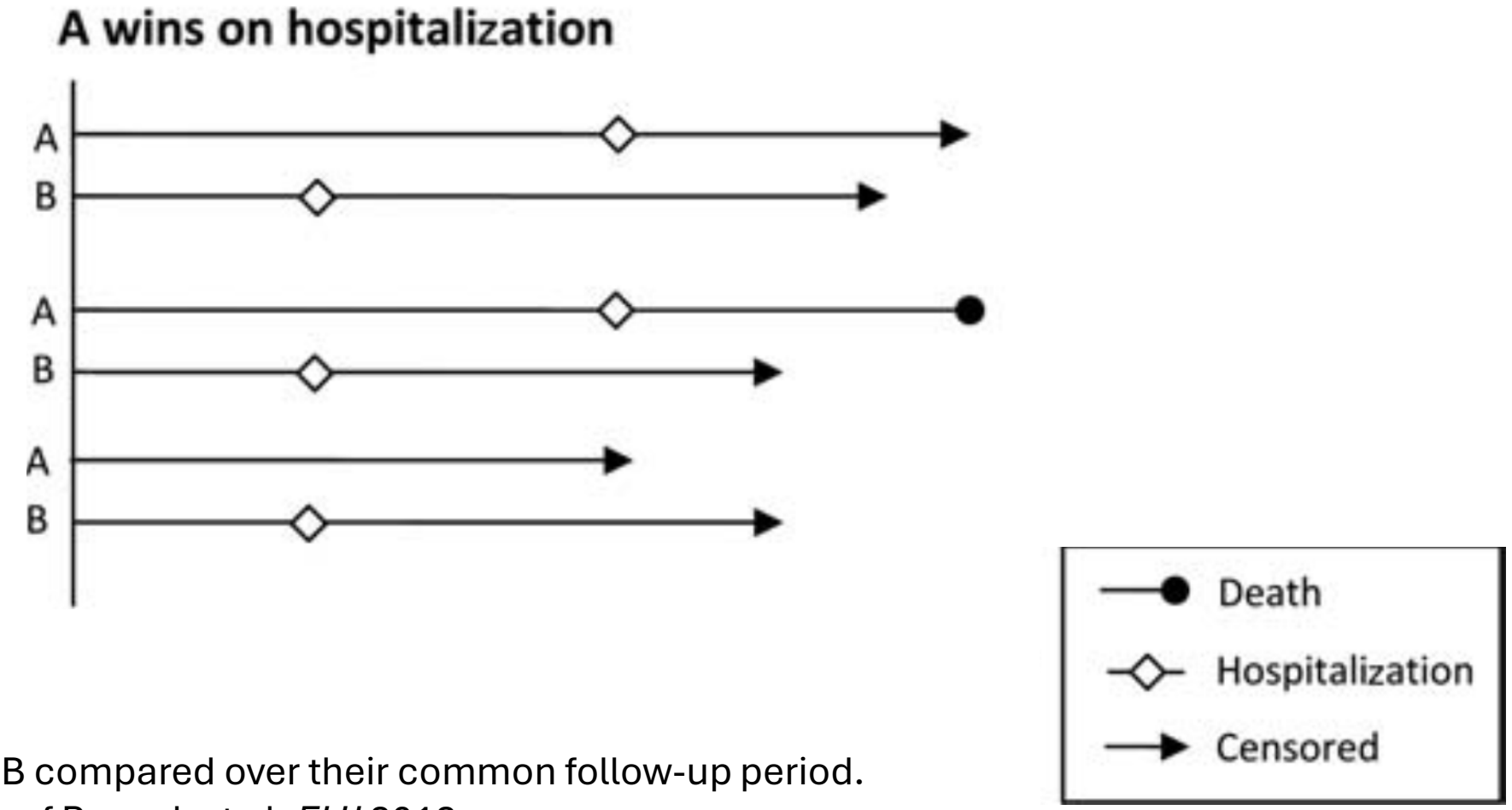
A wins on death





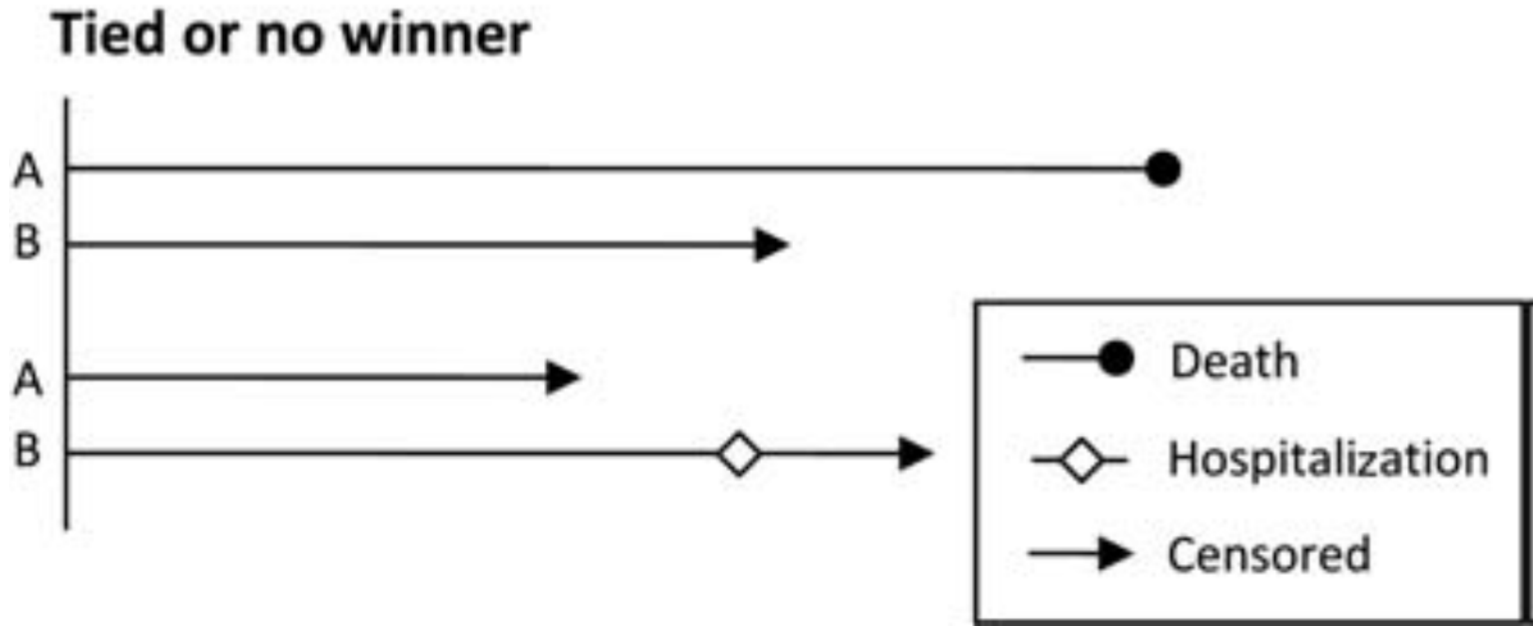
Patients A and B compared over their common follow-up period.
Figure courtesy of Pocock et al. *EHJ* 2012

A wins on hospitalization



Patients A and B compared over their common follow-up period.
Figure courtesy of Pocock et al. *EHJ* 2012

Winner cannot be determined



Patients A and B compared over their common follow-up period.
Figure courtesy of Pocock et al. *EHJ* 2012

HF-ACTION: 1060 × 1070 = 1,134,200 participant pairs for comparison

Component	Exercise wins	No Winner Determined	Control wins
1. Time to Death	144119 (12.7%)	863375 (76.1%)	126706 (11.2%)
2. Time to hospitalization	335697 (29.6%)	204799 (18.1%)	322879 (28.4%)
Overall	479816 (42.3%)	204799 (18.1%)	449585 (39.6%)

Cox model hazard ratio: Treatment hazard rate/control hazard rate: 0.93, 95% CI (0.83, 1.03), p = 0.14

1/hazard ratio = Control hazard rate/Treatment hazard rate: 1.08, 95% CI (0.98, 1.20), p = 0.14

$$\text{Win ratio} = \frac{\#\text{ Exercise wins}}{\#\text{ Control wins}} = \frac{479816}{449585} = 1.07, 95\%\ \text{CI}\ \ (0.96, 1.19), \text{p} = 0.25$$

$$\text{Win odds} = \frac{\#\text{ Exercise wins} + 0.5 \times \#\text{ ties}}{\#\text{ Control wins} + 0.5 \times \#\text{ ties}} = \frac{479816 + (0.5 \times 204799)}{449585 + (0.5 \times 204799)} = 1.05$$

$$\text{Proportion in favor of treatment} = \frac{\#\text{ Exercise wins} - \#\text{ Control wins}}{\text{total }\#\text{ comparisons}} = \frac{479816 - 449585}{1134200} = 2.4\%$$

P-values for win ratio, win odds, and proportion in favor of treatment are approximately the same.

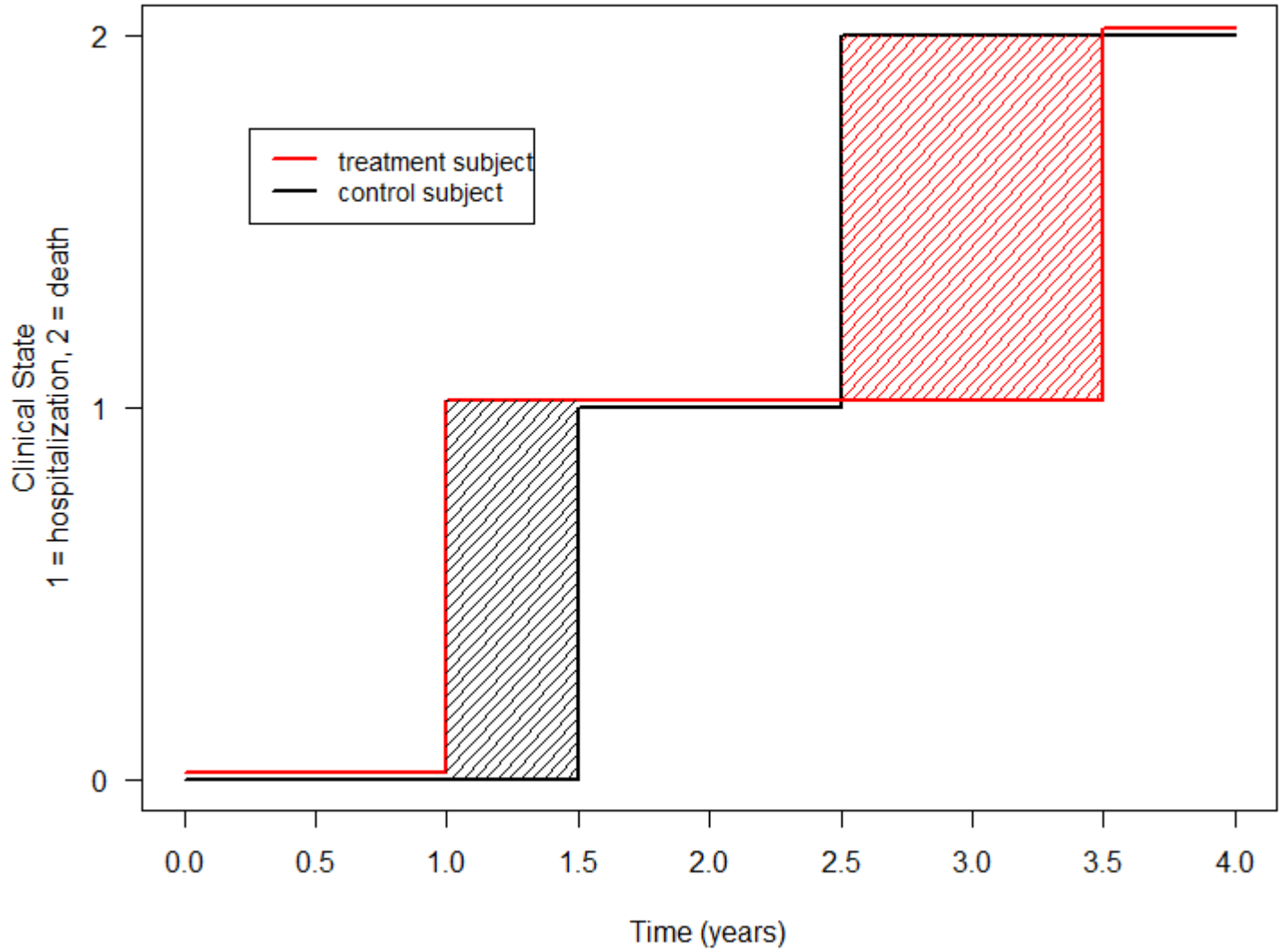
Win Time Ratio and Pairwise Win Time

- Like the Win Ratio, the Win Time Ratio and Pairwise Win Time compare all pairs of treatment-control subjects
- Win Time Difference for a particular pair equals the time the treatment subject is in a better clinical state MINUS the time the control subject is in a better clinical state

$$Win\ Time\ Ratio = \frac{\text{number treatment wins}}{\text{number of control wins}}$$

$$Pairwise\ Win\ Time = \text{average of the Win Time Differences}$$

Clinical state trajectories for two hypothetical patients

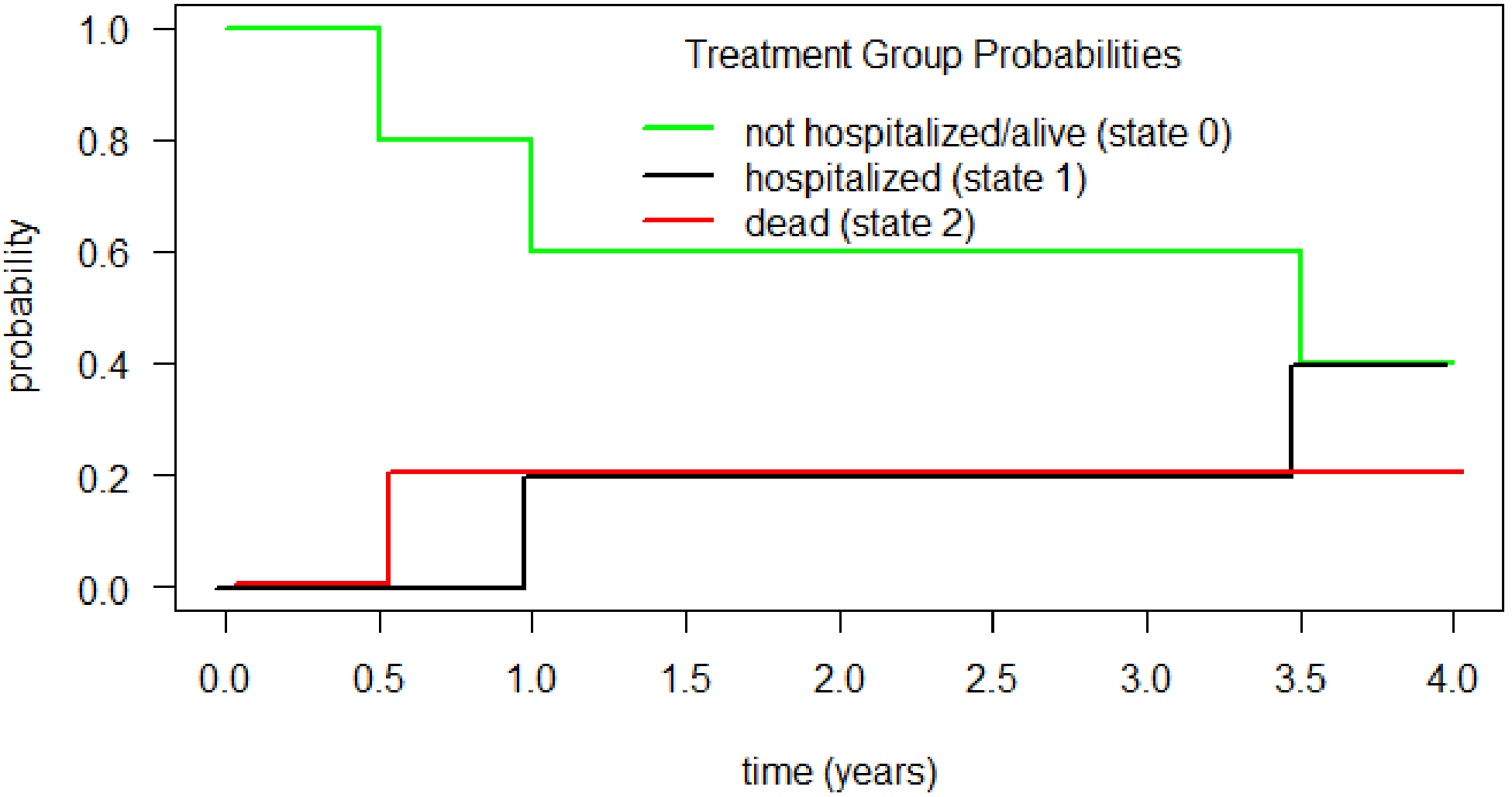


- Control subject is in a better clinical state for 0.5 years
- Treatment subject is in a better clinical state for 1 year
- Treatment wins this pair with 0.5 years spent in a better clinical state
- A treatment-control pair can be compared until the treatment or control subject is censored, whichever may happen first

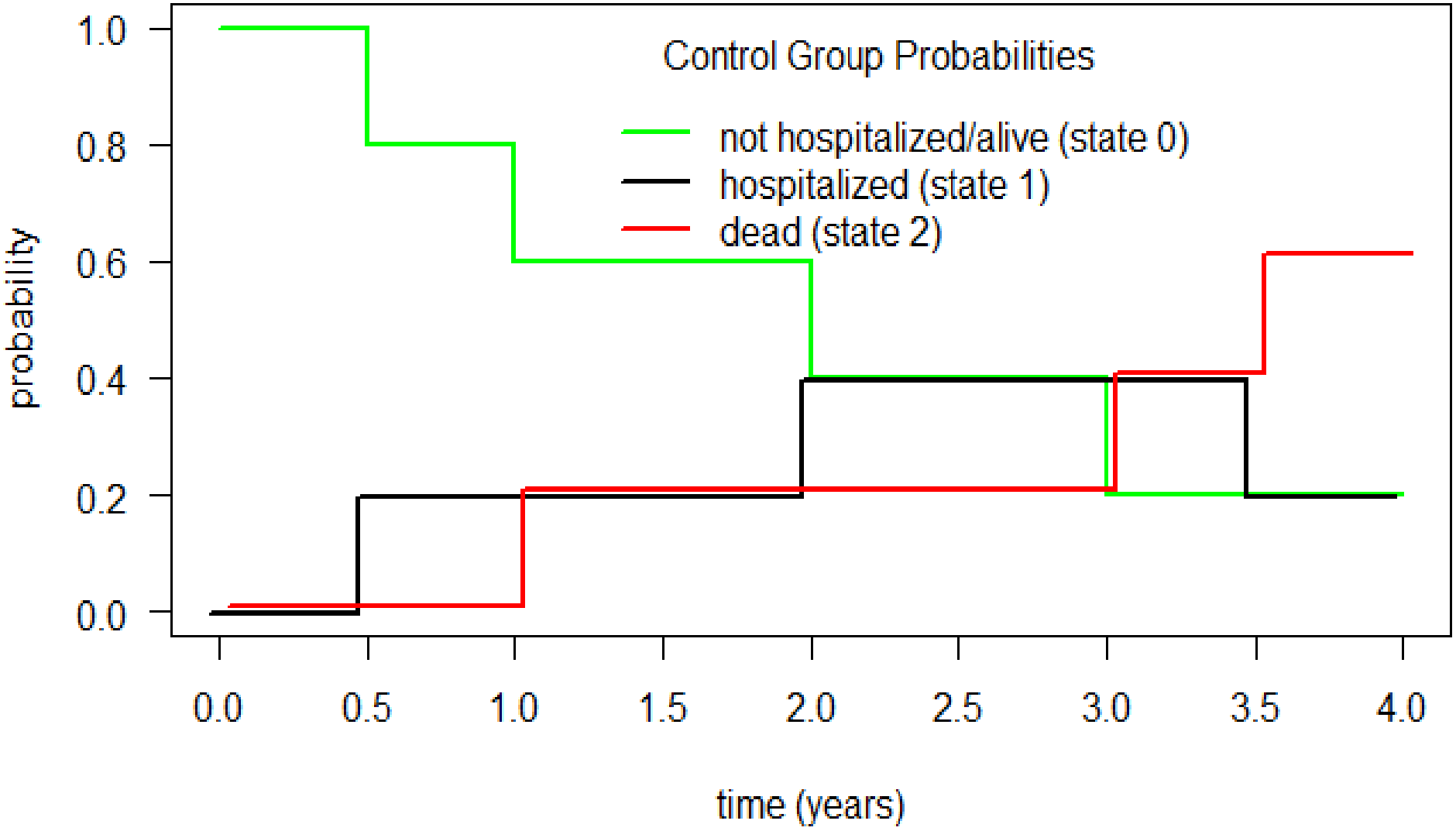
Expected Win Time

- No longer consider all treatment-control subject pairs.
- Instead, we use the treatment and control groups' probabilities of being in the different clinical states
- The average time a random treatment subject spends in a better clinical state than a random control subject MINUS the average time a random treatment subject spends in a worse clinical state than a random control subject.
- Equals the Restricted Mean Time in Favor of Treatment when there is a pre-specified follow-up time, e.g., 3 years

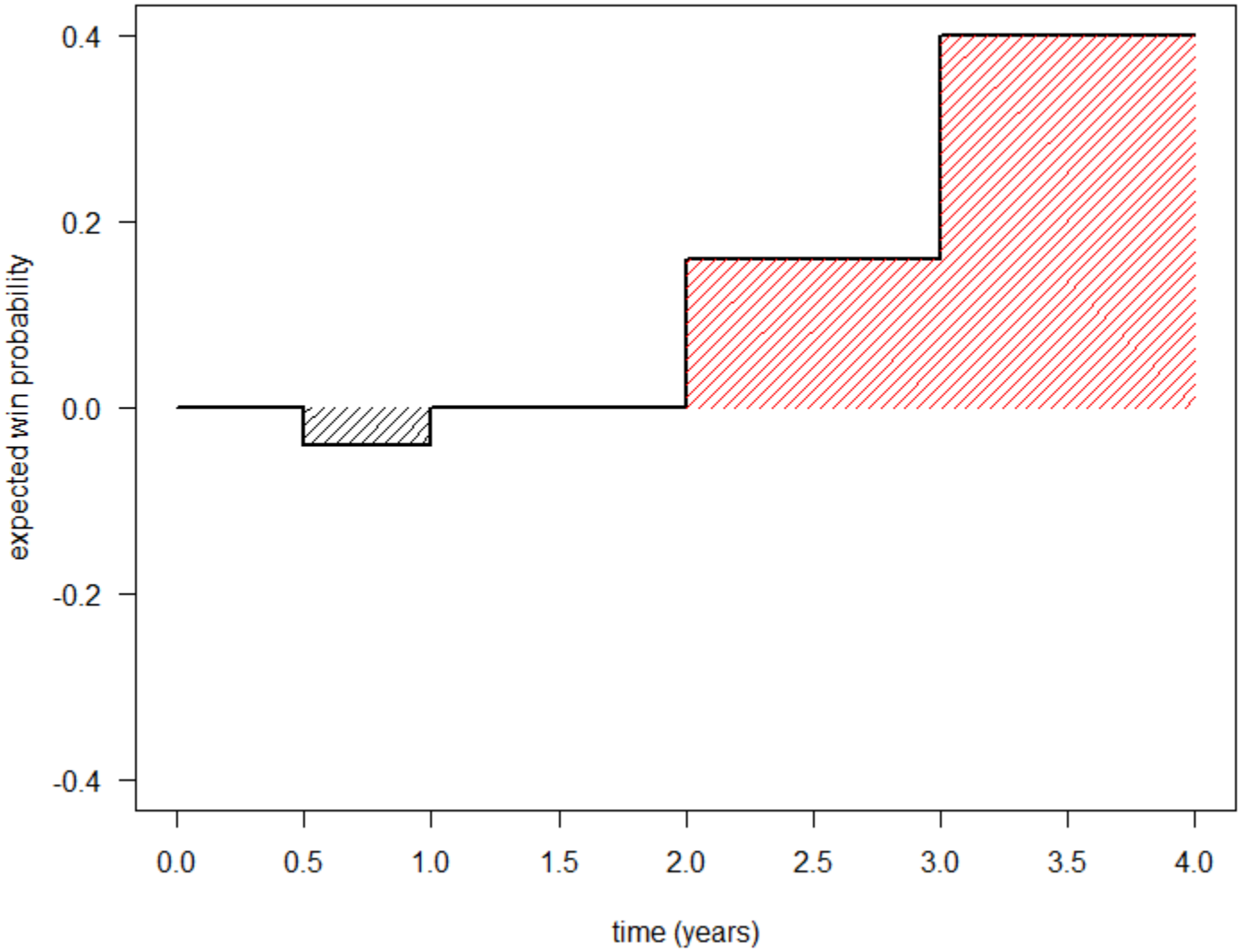
Hypothetical Treatment Group Clinical State Probabilities



Hypothetical Control Group Clinical State Probabilities



Expected Win Time



Clinical State Probabilities
Between Years 2 and 3

State	Treatment	Control
0 = no hosp/death	0.6	0.4
1 = hosp	0.2	0.4
2 = death	0.2	0.2

$$\text{Probability}(T > C) = 0.6 \times (0.4 + 0.2) + 0.2 \times (0.2) \quad = 0.40$$

$$\text{Probability}(C > T) = 0.4 \times (0.2 + 0.2) + 0.4 \times (0.2) = 0.24$$

$$\text{Expected Win Time between years 2 and 3} = (0.40 \ - 0.24) \times 1 = 0.16$$

$$\text{Expected Win Time} = \text{red area} \ - \text{black area} = (0.16 + 0.4) - 0.04 = 0.52 \text{ years}$$

Expected Win Time Against Reference

- Similar idea to the Expected Win Time except that instead of comparing the treatment group's clinical state probabilities to the control group's state probabilities, we compare each subject's clinical state trajectory to a reference group's (usually the control group) clinical state probabilities.
- Can adjust for prognostic covariates using a linear model and use linear regression for hypothesis testing

Statistical Power Comparison-set up

- 3-level hierarchy
  - death > heart failure > myocardial infarction
- 1000 simulated clinical trials
  - n = 2000 subjects
  - One-sided 0.025 significance level
  - Under null hypothesis, an average of:
    - 410 deaths
    - 764 heart failure events
    - 765 myocardial infarction events

Statistical Power Comparison - Results

	Death	HF	MI	Statistical Power (%)					
Scenario	Relative Risk Reduction			Cox Model	Win Ratio	Win Time Ratio	Pairwise Win Time	Expected Win Time by 3 years	Expected Win Time Against Reference
1	16%	16%	16%	**82**	78	78	71	69	**66**
2	10%	14%	18%	**78**	63	66	52	46	**38**
3	26%	18%	10%	**82**	91	88	92	**96**	94
4	30%	0%	–10%	**2**	24	14	51	62	**80**

HF-ACTION Results

	Cox Model 1/hazard ratio	Win Ratio	Win Time Ratio	Pairwise Win Time	Expected Win Time	Expected Win Time by 3 years	Expected Win Time Against Reference
Effect Estimate	1.08	1.07	1.07	22 days	44 days	32 days	37 days
P-value	0.14	0.25	0.20	0.14	0.058	0.054	0.10

Discussion (1)

- Each component in the hierarchy was a time-to-event
- Idea: more serious events should receive greater importance
- Win ratio and time-in-clinical state extensions can say if treatment had better outcomes with respect to the chosen hierarchy
- Choosing the hierarchy can be challenging

Discussion (2)

- Statistical power depends on treatment effect on the components
- Time-in-clinical state methods have higher power than a time-to-first event composite endpoint when there is a substantial treatment effect on death
- Need to separately look at the components to better understand the treatment effect

Acknowledgements

- James Troendle, Song Yang, Nancy Geller, Vandana Sachdev
- Office of Biostatistics Research, NHLBI, Colleagues
- Chris O’Connor, Mona Fiuzat, Mitch Psotka, Rob Mentz
- Brent Logan, Amber Fritz
- Valerie Leifer

References

Finkelstein DM, Schoenfeld DA. Combining mortality and longitudinal measures in clinical trials. *Statistics in Medicine*. 1999;18:1341-1354.

Buyse M. Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. *Statistics in Medicine*. 2010;29:3245-3257.

Pocock SJ, Ariti CA, Collier TJ, Wang D. The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. *European Heart Journal*. 2012;33:176-182.

Yang S, Troendle J. Event-specific win ratios and testing with terminal and non-terminal events. *Clinical Trials*. 2021;18:180-187.

Mao L. On restricted mean time in favor of treatment. *Biometrics*. 2023;79:61-72.

Troendle JF, Leifer ES, Yang S, Jeffries N, Kim DY, Joo J, O'Connor CM. Use of win time for ordered composite endpoints in clinical trials. *Statistics in Medicine*. 2024;43:1920-1932.

Leifer ES, Troendle JF, Psotka MA, Sachdev V. Hierarchical Analysis of Composite Time-to-Event End Points in Heart Failure Clinical Trials Using Time in Clinical State. *Circulation: Heart Failure*. 2025;18:e011783.

Extra Slides

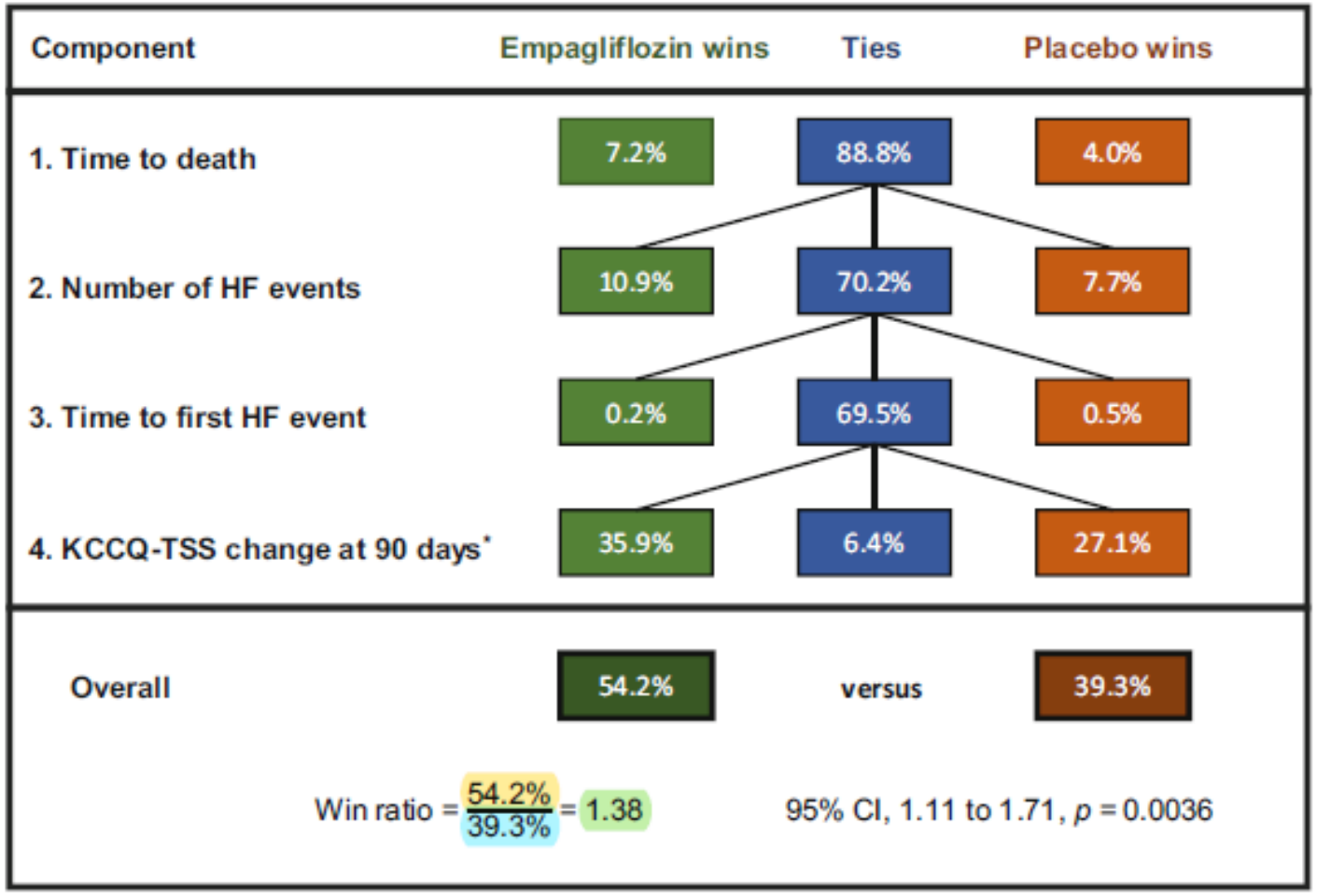
Discussion – what we covered

- We discussed various win ratio-type analyses
  - Each component in the hierarchy was a time-to-event
  - Idea: more serious events should receive greater importance
  - Statistical power depends on treatment effect on the components

Discussion – combining different endpoints

- Another reason to use a win ratio: combine different types of endpoints
- Finkelstein and Schoenfeld (1999)
  - Treatments for HIV
  - Combine mortality and longitudinal change in CD4 lymphocyte counts

Discussion – EMPULSE (Empagliflozin in Patients Hospitalized with Acute Heart Failure who have been Stabilized)



* note that the predefined primary analysis is stratified (see Figure 2 and Figure 3)
* a win requires at least 5 points difference between patients

265 × 265 paired comparisons

Cox model for time to death or HF event:
HR = 0.65 (0.43-0.99)

Courtesy of Pocock, et al. EJHF, 2023

Duke Clinical Research Institute

Recent Advances in Statistical Methods for Hierarchical Composite Endpoints

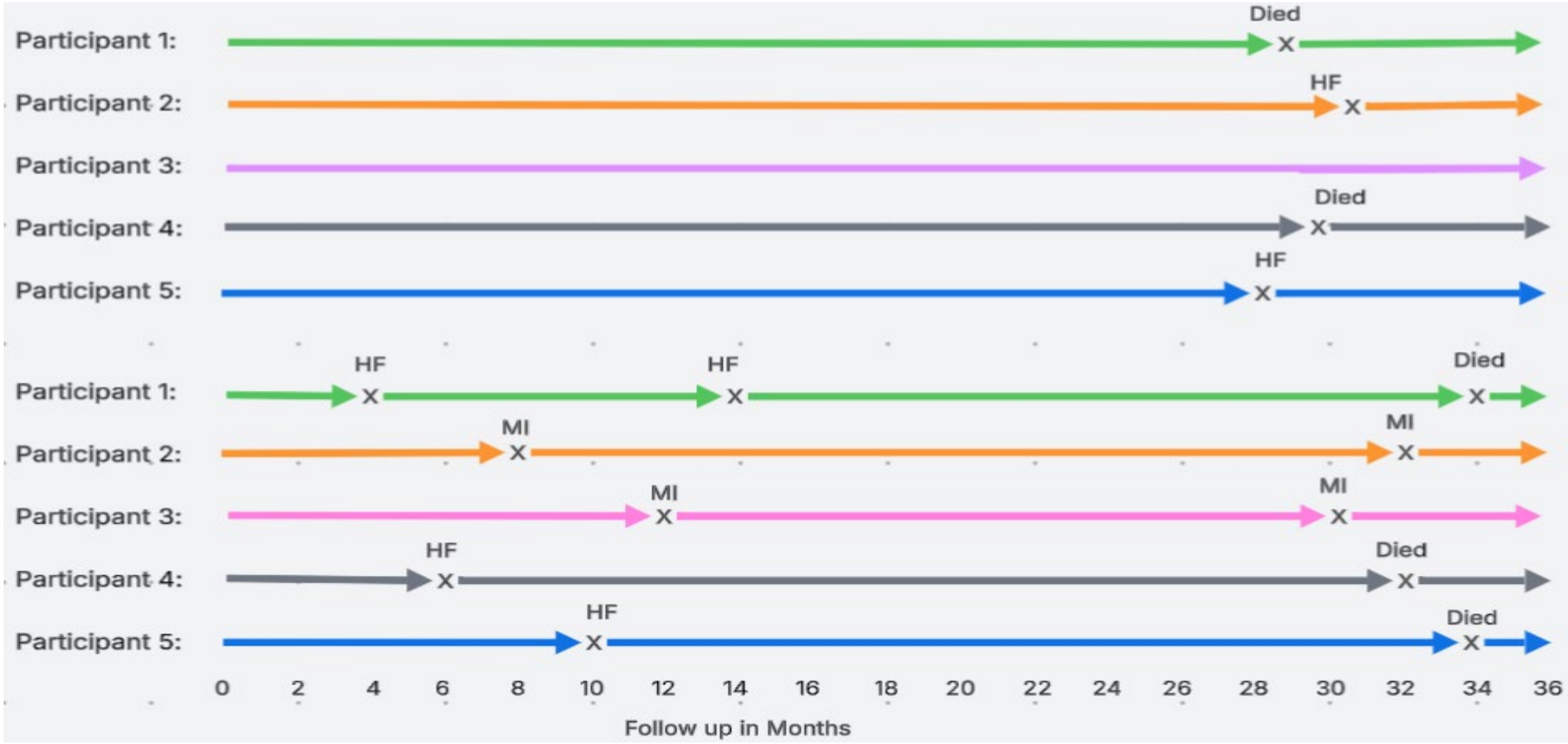
Roland A. Matsouaka, PhD
Duke University

Society for Clinical Trials Annual Meeting
May 18, 2026

Duke UNIVERSITY

No relevant disclosures

Beyond the standard win ratio



The 3 presentations

- Broader movement beyond time-to-first event (TTFE) analysis
  - Richer than TTFE, more clinically nuance than ordinary win ratio
- The methods bridge
  - classical win ratio methods,
  - multi-state/restricted-time ideas,
  - continuous patient-benefit summaries.
- Related methodological developments include:
  - DOOR/desirability methods,
  - restricted-time win ratio,
  - generalized pairwise comparisons,
  - hierarchical composite estimands.

Duke
UNIVERSITY

The usual issues and complications remain

- Censoring: drop out, treatment discontinuation, loss to follow up
- Missing longitudinal data
- Need to find best way to handle recurrent events
- Need for real trial applications, benchmarking vs. standard methods, consensus reporting guidance
- Formulas for power and sample size calculation

Duke
UNIVERSITY

Lu Mao

WRNet: Regularized win ration regression

Core idea:

- Extends WR method into high-dimensional regression to allow
  - Variable selection, shrinkage, and risk prediction
- Generalizes Cox PH regression to HCE
- Enables the use of LASSO, ridge, and elastic net penalties
- Provides an R package to implement *wrnet*

Duke
UNIVERSITY

Limitations

- Only for time-to-event endpoints
- Assumption: Proportional hazards over each covariate

Questions:

1. Can we use the win indicator to extend to mixed types of endpoints?
2. How can the PH assumption be weakened?

Duke
UNIVERSITY

Eric Leifer

Extending the win ratio to time spent in better clinical states

Core idea:

Move beyond “who wins?”

to

“How long do patients remain in better clinical states?”

The paper’s message: Hierarchical composite endpoint analyses should measure not only *who wins*, but *for how long patients remain better off*.

Duke
UNIVERSITY

Advantages and limitations

Advantages:

- More patient-centered and uses more information
- Reflects “time feeling/doing better,” not just event occurrence.
- Incorporates longitudinal patient experience
- Compares time spent in superior states

Limitations:

- Methods apply to time-to-events only
- Interpretation: clinical meaning of additional “win-days”?
- Implicit time weighting and may not reflect clinical value
  - E.g.: 3 extra months alive with severe disability vs. shorter survival with excellent QoL
- Monotonicity

Question:

- Weight time in clinical states by QoL, symptom burden, patient preference?

Duke UNIVERSITY

Huiman Barnhart

Weighted composite endpoints approach: an alternative to win ratio

Core idea:

Not all endpoints are equally important

Solution:

Assigns severity weights and aggregates endpoints into a score

- E.g., death > myocardial infarction > hospitalization

Duke UNIVERSITY

Advantages and limitations

- Advantages
  - Reflect clinical importance and emphasize clinically meaningful outcomes
  - Better capture total disease burden over follow up
  - Align analysis with how clinicians and patients value outcomes
- Limitations
  - Results may depend heavily on selected weights
  - Potential disagreement among clinicians, patients, sponsors, regulators, …
  - Interpretation: clinical meaning of the effect size?
- Question
  - How a sensitivity analysis based on weights can look like?

Duke
UNIVERSITY

Q & A

Duke
UNIVERSITY