From Composite Outcomes to Win Ratio: Applications, Extensions, and Limitations

Rejuan Haque

The Ohio State University Wexner Medical Center
47th Annual Meeting of the Society for Clinical Trials

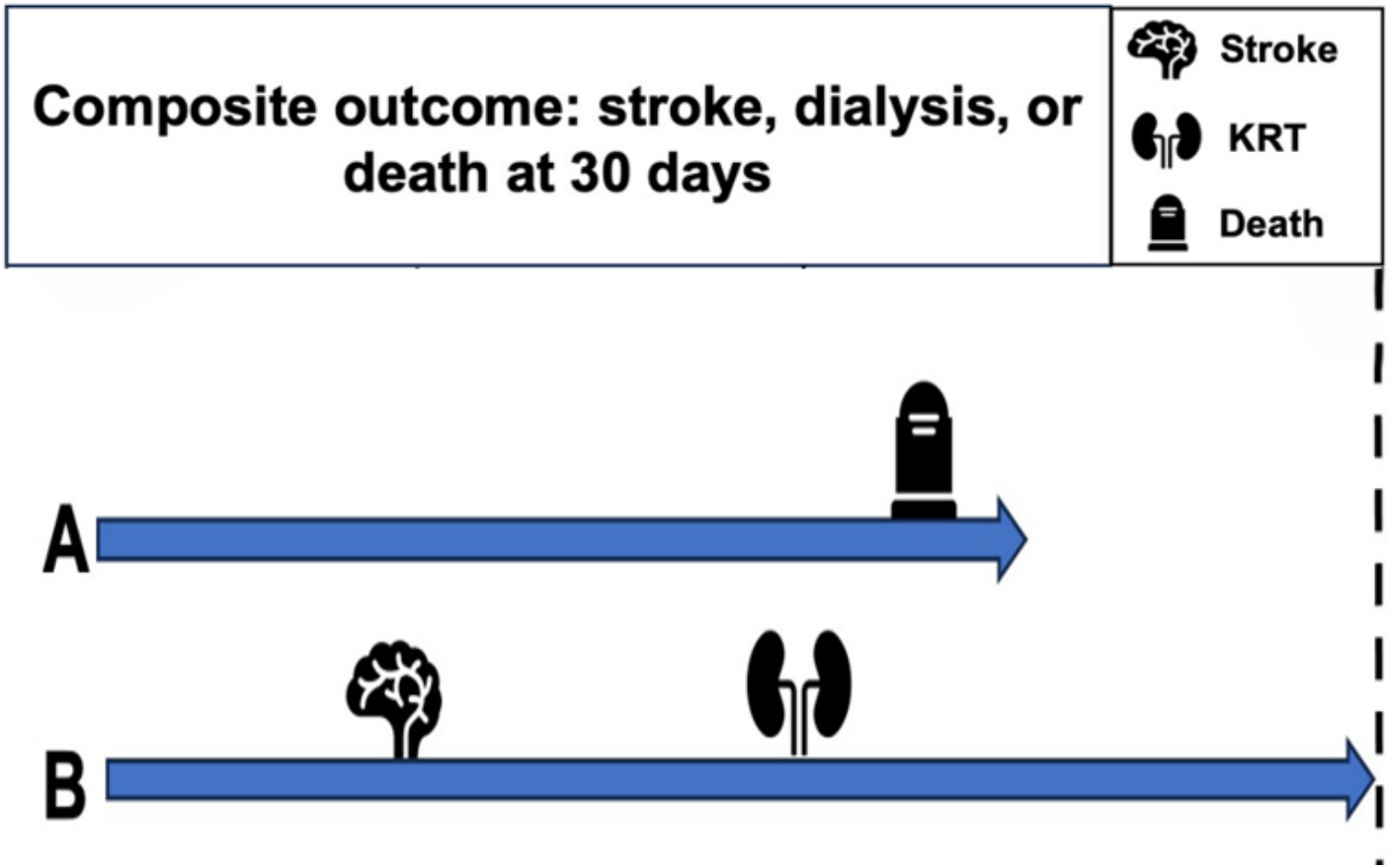
What is Composite Outcome?

- Composite endpoints
  - Death and cardiovascular (CV) hospitalizations due to heart failure, stroke, or myocardial infarction
  - Death, ICU length of stay, Hospital length of stay
- Traditional methods
  - Time to first event - univariate method (e.g., Kaplan-Meier curves, log-rank test, Cox model)
  - Logistic regression model
- Limitations:
  - Ignores later events (e.g., death after progression)
  - Treats all events equally (e.g., Death = hospitalization).
  - Not suitable for mixed-type outcomes
  - Statistical power and sample size consideration

Win Ratio: A Better Approach

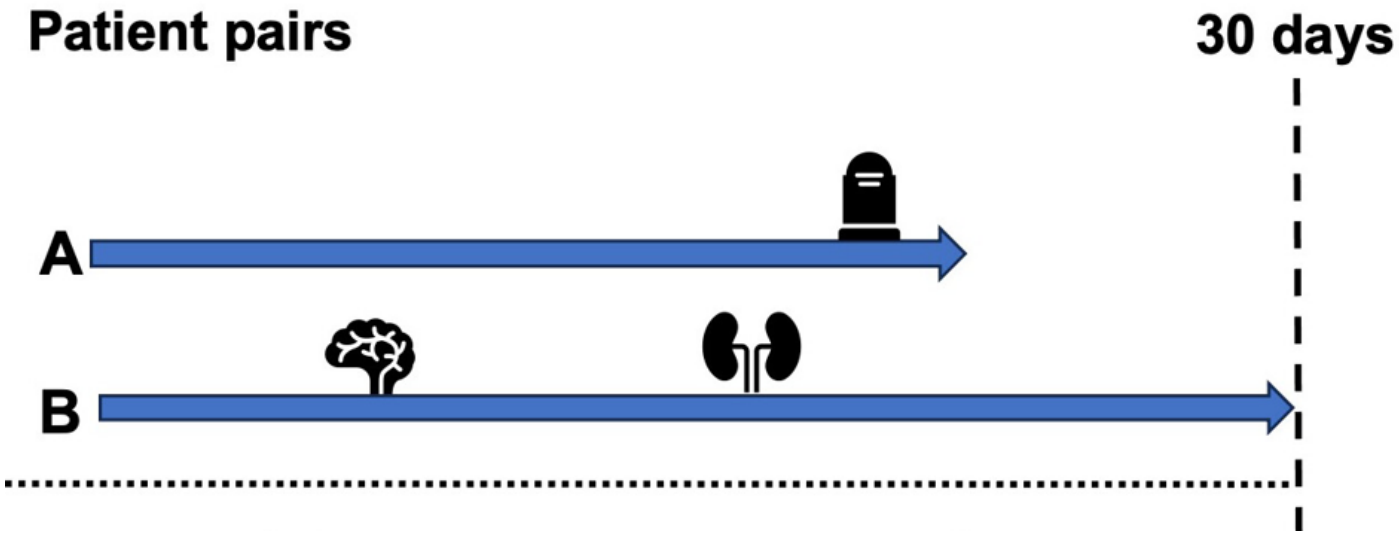
- Proposed by Pocock et al. (2012)
- Core Principle: Pairwise comparisons using clinical priority hierarchy (e.g., Death $>$ HF hospitalization $>$ Urgent visit)
- Forms every possible patient-to-patient pair
  - Win: The patient on the new treatment has a better outcome
  - Loss: The control patient does better
  - Tie: Neither win nor loss
- Win or Loss is decided based on a pre-defined Win Rule using the hierarchy of the outcomes (from most to least important)

Who wins?



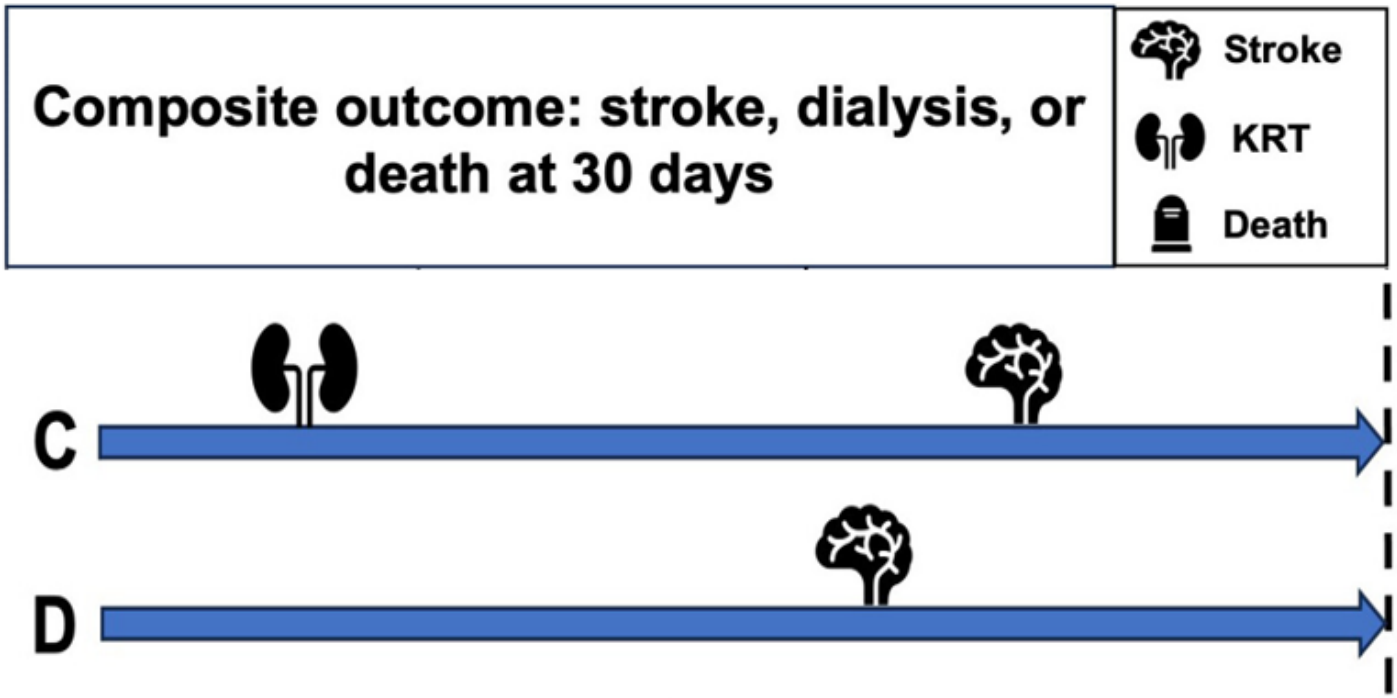
- Hierarchy: Death $>$ Stroke $>$ KRT

Who wins?



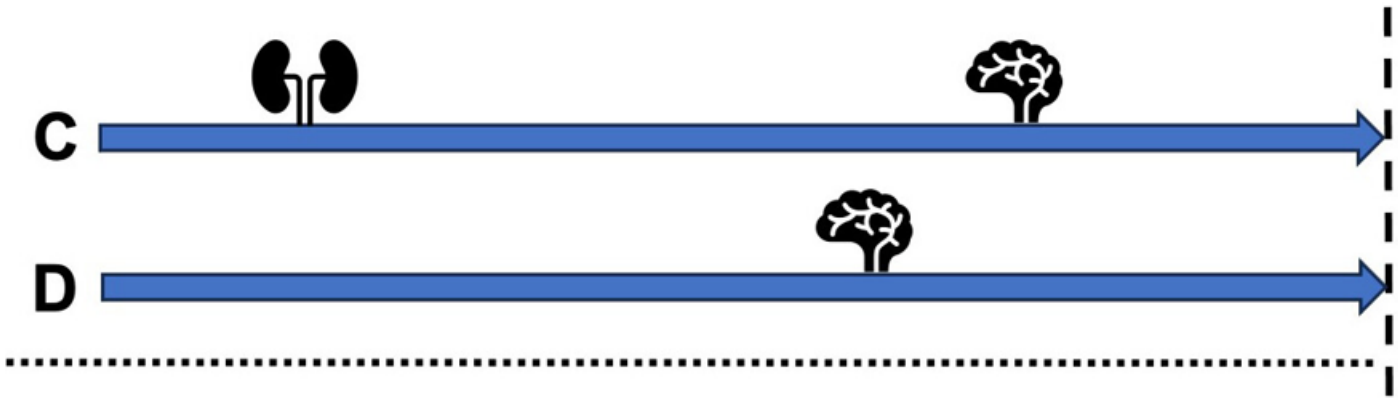
Logistic regression **"Tie"** *(both patients had an event of composite at 30d)*
Cox analysis **Patient A "wins"** *(first event of composite later than patient B)*
Win ratio (*Hierarchy: Death > Stroke > KRT*) **Patient B "wins"** *(1^{st} event in hierarchy [death] at 30d only in patient A)*

Who wins?



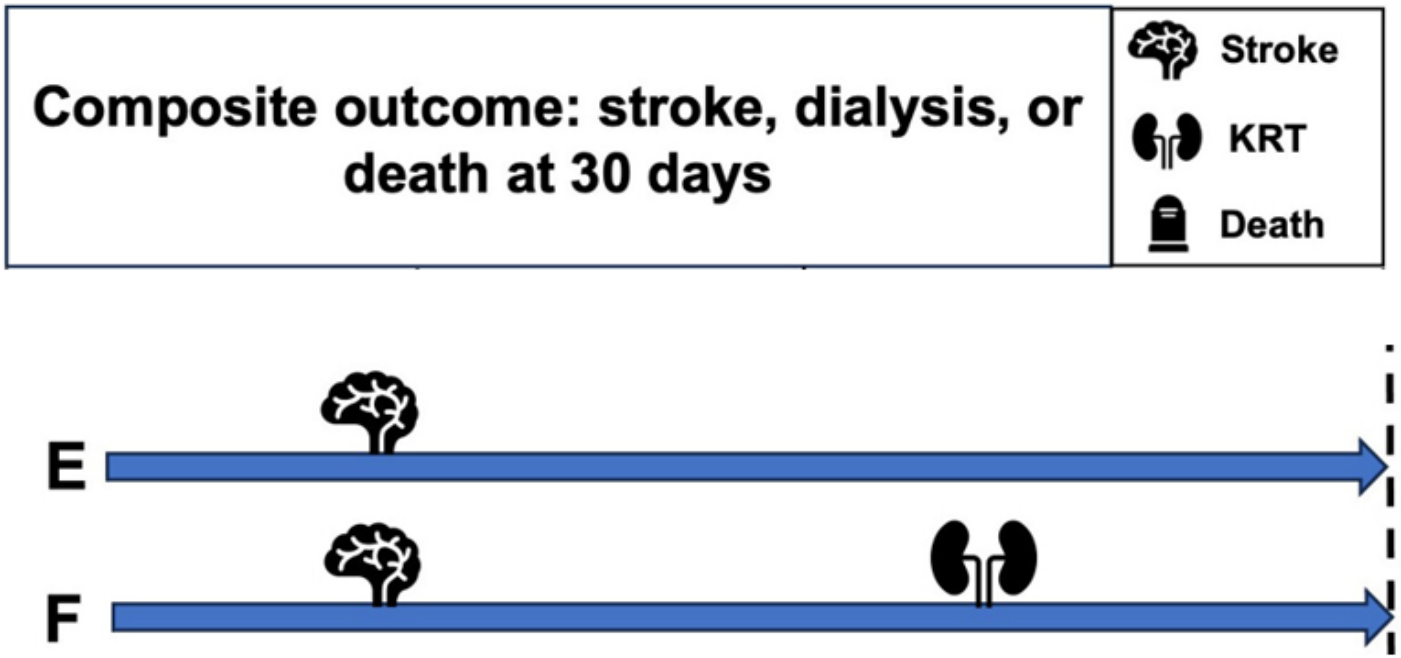
- Hierarchy: Death $>$ Stroke $>$ KRT

Who wins?



Logistic regression **"Tie"** *(both patients had an event of composite at 30d)*
Cox analysis **Patient D "wins"** *(first event of composite later than patient C)*
Win ratio (*Hierarchy: Death > Stroke > KRT*) **Patient D "wins"** *(Pair tied for 1st [death – no event] and 2nd [stroke – both] event in hierarchy at 30d, KRT only in patient C)*

Who wins?

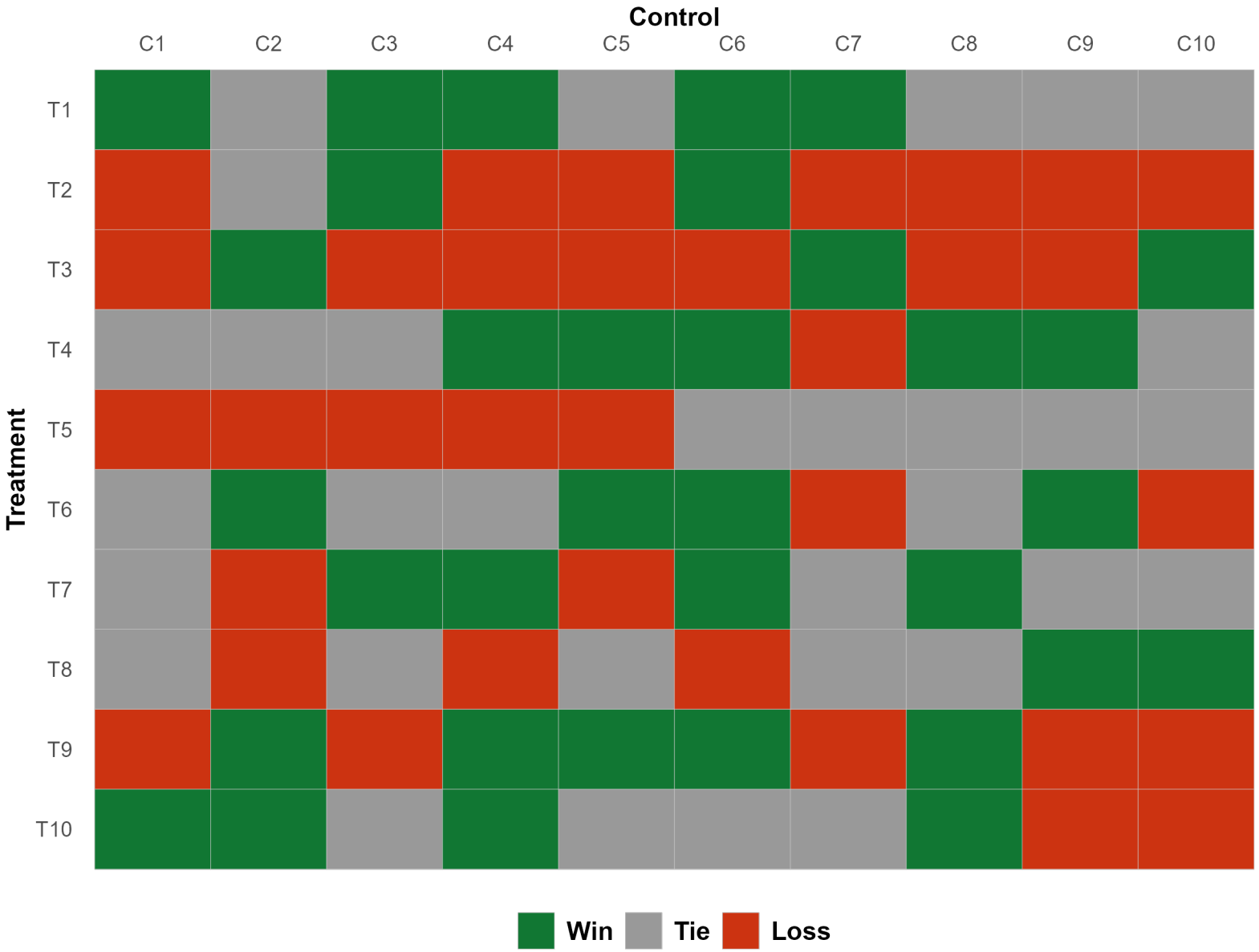


- Hierarchy: Death $>$ Stroke $>$ KRT

Who wins?

Logistic regression **"Tie"** *(both patients had an event of composite at 30d)*
Cox analysis **"Tie"** *(first event of composite [stroke] same time)*
Win ratio (*Hierarchy: Death > Stroke > KRT*) **Patient E "wins"** *(Pair tied for 1st [death] and 2nd [stroke] event in hierarchy at 30d, KRT only in patient F)*

Win matrix


Control
C1
C2
C3
C4
C5
C6
C7
C8
C9
C10
Treatment
T1
T2
T3
T4
T5
T6
T7
T8
T9
T10
Win
Tie
Loss

Win ratio (WR)

- Let n and m be the number of patients in the Treatment and Control groups
- Make $n \times m$ paired comparison
- Let n_w and n_l be the number of wins and losses out of the $n \times m$ paired comparison
- The win ratio WR $= \frac{n_w}{n_l}$

Statistical Inference for the Win Ratio

- Hypothesis Test

$$H_0 : WR = 1 \quad [Pr(Win) = Pr(Loss)]$$
$$H_1 : WR \neq 1 \quad (Two\text{-}sided)$$

- Test Statistic

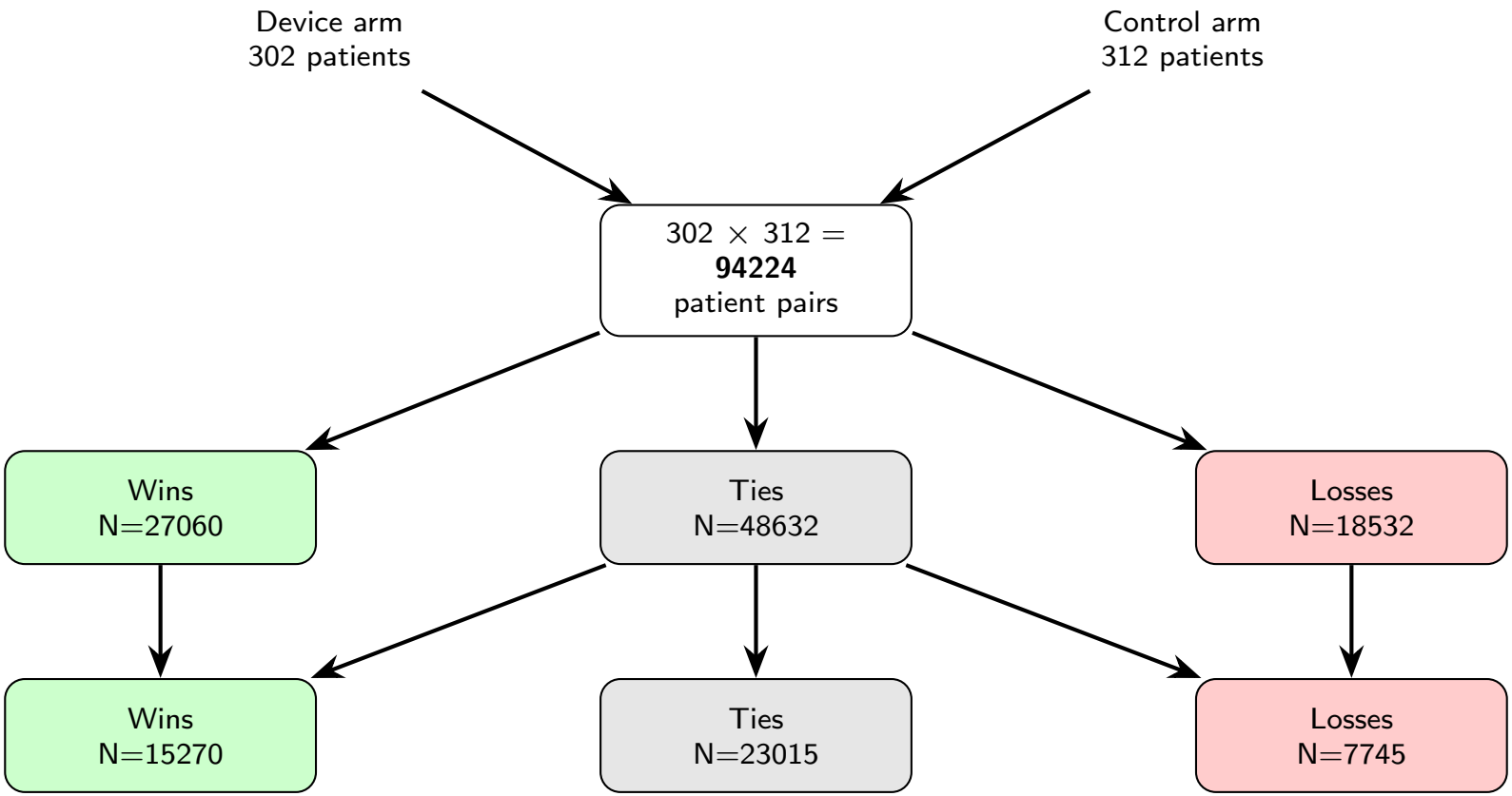
$$Z = \frac{\ln(WR)}{SE(\ln(WR))} \sim N(0,1) \text{ under } H_0 \text{ (Bebu et al., 2016)}$$

- 95% Confidence Interval: $Exp[\log(WR) \pm 1.96 \times SE]$
  - Bebu & Lachin (2016) provided the exact U-statistic-based variance, valid for small samples to calculate the test statistic and CI for the Win Ratio

Example

- The EMPHASIS-HF trial (Pocock et al. 2012)
- Composite endpoint of death and the number of hospitalizations due to heart failure.
- Hierarchy:
  - Tier 1: Death
  - Tier 2: Heart failure hospitalizations

Example



Win ratio = total wins / total losses
= 42330 / 26277 = **1.61**
95% CI (1.29, 2.04), $p < 0.0001$

Applications: Cardiovascular Clinical Trials

- **ATTR-ACT Trial**
  - Transthyretin amyloidosis (Tafamidis)
  - Hierarchy: All-cause death $\rightarrow$ CV hospitalization
- **PARTNER B Trial (TAVI)**
  - Inoperable aortic stenosis
  - Hierarchy: Death $\rightarrow$ Hospitalization
- **EMPHASIS-HF Trial**
  - Heart failure (Eplerenone)
  - Hierarchy: CV death $\rightarrow$ HF hospitalization
- **CHARM Program**
  - Candesartan in heart failure
  - Re-analysis using Win Ratio
- **EMPULSE Trial**
  - Acute heart failure (Empagliflozin)
  - Death $\rightarrow$ HF events $\rightarrow$ Symptoms

Applications Beyond Cardiovascular Trials

- **Infectious Disease / COVID-19**
  - ACTION Trial (Anticoagulation in COVID-19)
- **Critical Care / Respiratory**
  - Ventilator-Associated Pneumonia Trial (Mortality $\rightarrow$ Ventilator-free days)
- **Pulmonary Hypertension**
  - INCREASE Trial (Clinical worsening $\rightarrow$ Functional outcomes)
- **Neurology**
  - Stroke and disability outcomes (emerging use of hierarchical comparisons)
- **Oncology**
  - Composite endpoints: Death $\rightarrow$ Progression $\rightarrow$ Toxicity
- **Surgery / Perioperative**
  - Hepatopancreatic surgery (Mortality $\rightarrow$ Complications $\rightarrow$ Readmission)

NIH-Funded and Active Win Ratio–Related Studies

Win Ratio is increasingly used in active NIH-supported clinical research programs:

- DEPRESCRIBE-HFpEF — HFpEF deprescribing strategy trial
- WINDSURFER — Emergency department respiratory failure trial
- CORD-CHD — Neonatal congenital heart disease (cord clamping strategies)
- MAPT — Musculoskeletal adaptive platform trial

Growing adoption of Win Ratio in pragmatic and platform trial designs

Methodological Advances in Win Ratio

- **Statistical Theory**
  - Bebu et al. (2016); Dong et al. (2016): Asymptotic distribution using U-statistic theory
  - Variance estimation and inference framework
- **Design Extensions**
  - Dong et al. (2018); Gasparyan et al. (2021) : Stratified Win Ratio: accounts for baseline heterogeneity
- **Weighted Approaches**
  - Luo et al. (2017); Haque et al. (2026): Weighted Win Ratio: incorporates covariate-based or clinical weighting
- **Regression Frameworks**
  - Mao et al. (2021): Win Fraction Regression (Proportional Win Model)
  - Allows covariate adjustment and individualized treatment effect estimation
- **Time-to-Event and Censoring**
  - Dong et al. (2020): IPCW-based Win Ratio for censored outcomes
  - Extensions to survival data and recurrent events

Software for Win Ratio Analysis

- **R Packages**
  - `WinRatio`: Basic implementation of win ratio and stratified WR
  - `WR`: Functions for matched and unmatched win ratio analyses
  - `WINS`: Advanced tools for hierarchical composite endpoints
- **Stata**
  - `winratio`: Command for computing win ratio and confidence intervals
  - Supports stratified analyses and hypothesis testing
- **Other Implementations**
  - Custom code (R / Python) for:
    - Weighted Win Ratio
    - IPCW-based methods
    - Win fraction regression

Most advanced extensions (e.g., weighting, regression) currently require custom implementation

Extensions Beyond the Win Ratio

- **Win Odds**
  - Incorporates ties into the estimand
  - Defined as:
$$\frac{P(\text{Win}) + 0.5P(\text{Tie})}{P(\text{Loss}) + 0.5P(\text{Tie})}$$
  - More stable when ties are frequent
- **Net Benefit**
  - Direct probability contrast:
$$P(\text{Win}) - P(\text{Loss})$$
  - Symmetric and interpretable estimand
  - Avoids ratio-based interpretation

Potential Pitfalls of the Win Ratio

- **Treatment of ties**
  - Ties are excluded from the Win Ratio
  - Can lead to loss of information when ties are common
- **Dependence on outcome hierarchy**
  - Results can change with different ordering of endpoints
  - Requires strong clinical justification
- **Interpretability**
  - Ratio scale (Wins/Losses) may be less intuitive than probabilities or odds
- **Computational burden**
  - Requires all pairwise comparisons
  - Can be intensive in large trials
- **Censoring and survival assumptions**
  - Requires additional methods (e.g., IPCW) for valid inference

Conclusion

Key takeaways on the Win Ratio framework:

- The Win Ratio provides a **clinically meaningful framework** for analyzing hierarchical composite outcomes
- It improves upon traditional composite endpoints by **respecting outcome priorities**
- The method has expanded from a simple summary measure to a **rich inferential framework** (e.g., stratified, weighted, regression-based, IPCW extensions)
- It is now widely used across **cardiology, critical care, oncology, neurology, and pragmatic trials**
- Limitations such as **ties, hierarchy dependence, and interpretation challenges** have motivated alternatives like **Win Odds and Net Benefit**

Key References

Pocock SJ, et al. (2012). *The win ratio: a new approach to the analysis of composite endpoints in clinical trials.* **Stat Med.**

Bebu I, Lachin JM (2016). *Large sample inference for a win ratio analysis of a composite outcome based on prioritized components.* **Biostatistics, 17(1):178–187.**

Dong G, Li D, Ballerstedt S, Vandemeulebroecke M (2016). *A generalized analytic solution to the win ratio.* **Pharm Stat, 15(5):430–437.**

Dong G, Qiu J, Wang D, Vandemeulebroecke M (2018). *The stratified win ratio.* **J Biopharm Stat, 28(4):778–796.**

Luo X, Qiu J, Bai S, Tian H (2017). *Weighted win loss approach for analyzing prioritized outcomes.* **Stat Med, 36(15):2452–2465.**

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

Thank You!

Questions or Comments?

Rejuan Haque
mdrejuan.haque@osumc.edu

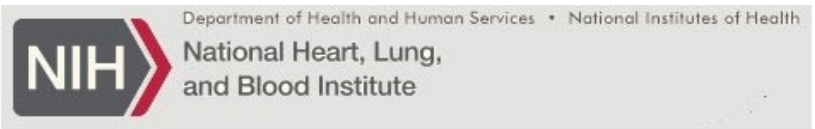
Extensions of Win Time Methods for Clinical Trials

James F. Troendle

Office of Biostatistics Research

Division of Intramural Research
NHLBI, NIH

SCT Meeting
May 19, 2026



Disclosures

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), and NHLBI (National Heart, Lung, and Blood Institute).

The contributions of the NIH authors are considered Works of the United States Government. The findings and conclusions presented in this talk are those of the authors and do not necessarily reflect the views of the NIH or the U.S. Department of Health and Human Services.

Presentation Outline

1. Clinical Trial Context
2. Hierarchical Methods
3. Win Time Methods
4. HF-ACTION
5. Conclusions

Clinical Trial Context

Composite Event

- Primary Outcome: Time to first of several clinical events including death
  - Usually log-rank test used or a Cox model with treatment term
  - Non-fatal events included to increase power if disease is not highly fatal
- HF-ACTION (O'Connor et al. 2009)
  - Trial of Exercise for HF-subjects
  - Sample size of 2331 randomized to exercise training or usual care (control)
  - Primary outcome was composite of death and hospitalization

Major Issue with Interpretation of Composite Event Results

- If treatment provides benefit on hospitalization and harm on death, it could look good based on the composite but might not provide overall benefit: what conclusion for use of treatment?

Win Ratio

- Use pairwise comparisons according to a hierarchy
- Every pair compared on death and if inconclusive then compared on hospitalization, etc.
  - Finkelstein and Schoenfeld (1999)
  - Pocock et al. (2012)

Win Ratio

Example

- TRILUMINATE (Sorajja et al. 2023)
  - Trial of TEER therapy for patients with tricuspid regurgitation
  - Sample size of 350 randomized to TEER or medical therapy (control)
  - Primary outcome win ratio based on
    1) death or tricuspid valve surgery
    2) # of hospitalizations for Heart Failure
    3) $\geq$ 15 point improvement from baseline in Kansas City Cardiomyopathy Questionairre (KCCQ)

TEER: Trans-catheter edge-to-edge repair

TRILUMINATE: 175 × 175 = 30625 participant pairs for comparison

Component	TEER wins	Ties	Control wins
1. Time to Death or TV surgery	2884 (9.4%)	25097 (81.9%)	2644 (8.6%)
2. Number of HF hospitalizations	1948 (6.4%)	20278 (66.2%)	2871 (9.4%)
3. KCCQ change	6516 (21.3%)	11634 (38.0%)	2128 (6.9%)
Overall	11348 (37.1%)	11634 (38.0%)	7643 (25.0%)

Primary analysis: Finkelstein-Schoenfeld = $\frac{\#\text{ TEER wins}}{\text{standard deviation}} = \frac{11348}{4891} = 2.32$

Win ratio = $\frac{\#\text{ TEER wins}}{\#\text{ TEER losses}} \quad \frac{11348}{7643} = 1.48$

Win odds = $\frac{\#\text{ TEER wins}+0.5}{\#\text{ TEER losses}+0.5} = \frac{11348+(0.5 \times 11634)}{7643+(0.5 \times 11634)} = 1.28$

Proportion in favor of treatment = $\frac{\#\text{ TEER wins}-\#\text{ TEER losses}}{\text{total \# comparisons}} = \frac{11348-7643}{30625} = 0.12$

All methods have approx. p-value = 0.0204

Win Ratio Issues

- Prioritizes most important events better than composite event method
- Does not remove major concern that it can be driven by least important events (as in TRILUMINATE)
- Win Odds or (excess) proportion in favor of treatment better reflect treatment effect when there are many ties
- Estimand depends on censoring

Multi-state Models

- (1) Use multiple Kaplan-Meier (KM) curves to infer distribution at any time
- (2) Use continuous-time Markov model to get transition probabilities at any time
- Markov method requires assumption, but transition probabilities provide additional value

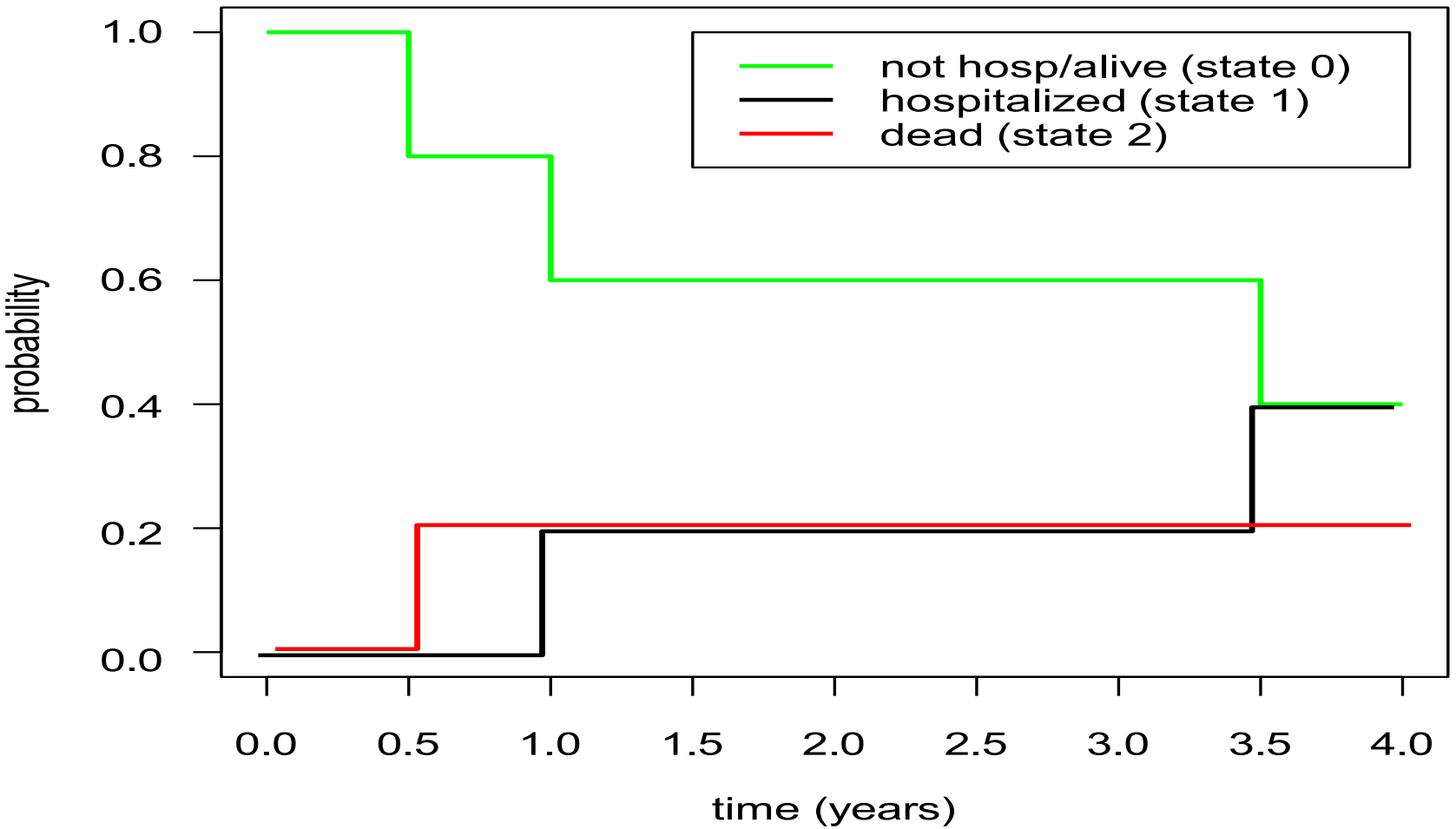


Figure: Combined Arm Probabilities

Expected Win Time Against Population (EWTP)

- Calculate the State probabilities for combined arms (trial population)
- Calculate an Expected Win Function (EWF) for each subject
- Integrate EWF to get EWTP value for each subject
- Use linear model of the EWTP values to get difference between arms
- Treatment Effect: treatment leads to an average of (for example) 0.5 excess years in a better state than in a worse state for treatment subjects compared to control subjects over the average follow-up time

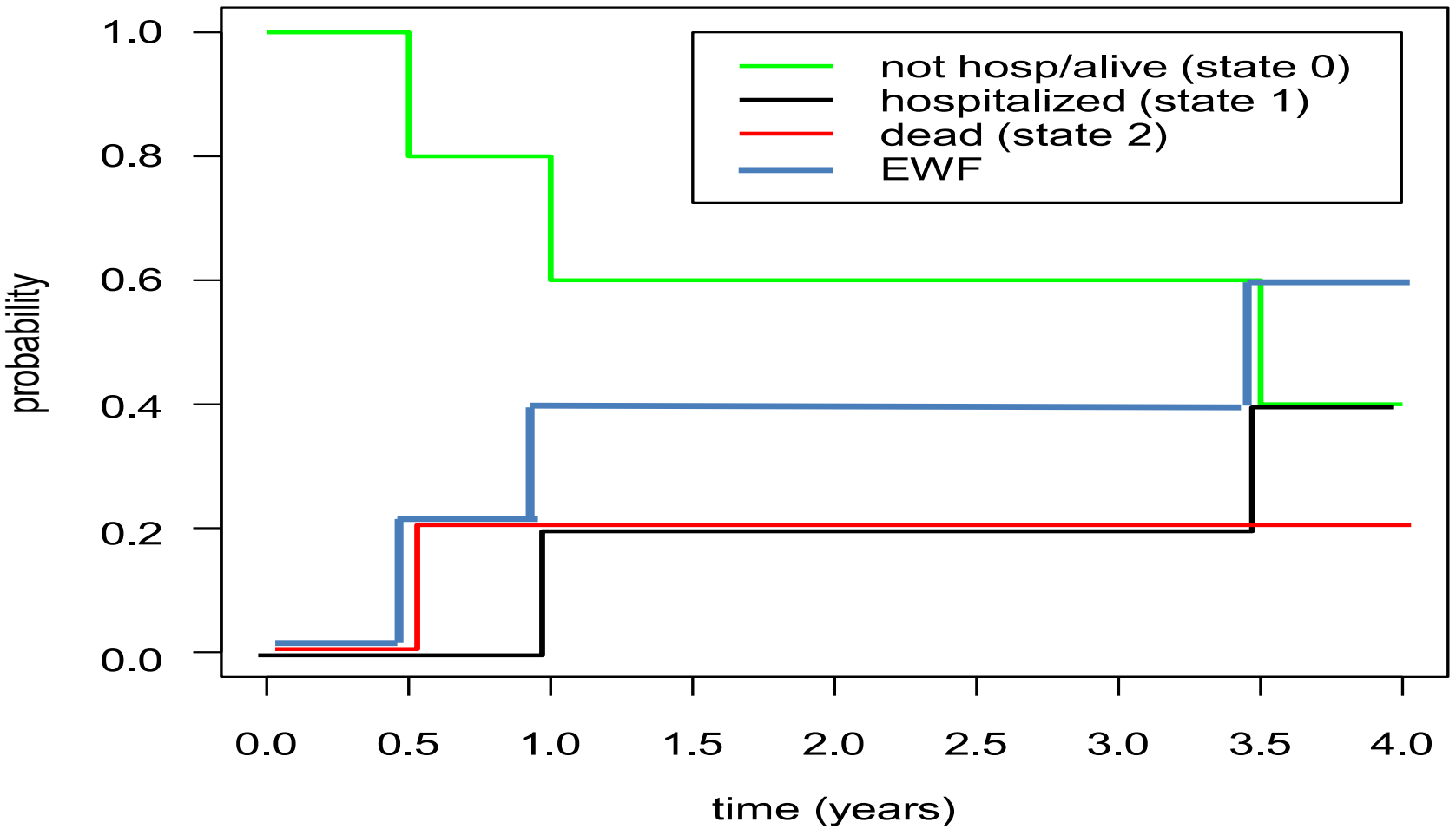


Figure: Combined Arm Probabilities and EWF for Subject Alive for 4 years

Expected Win Time Against Population With Redistribution (EWTPR)

- Calculate the State probabilities for combined arms (trial population)
- If censored, extend each person's follow-up by imputing future states using transition probabilities from Markov model (multiple imputation used)
- Calculate an Expected Win Function (EWF) for each subject
- Integrate EWF to get EWTPR value for each subject
- Use linear model of the EWTPR values to get difference between arms
- Treatment Effect: treatment leads to an average of (for example) 0.5 excess years in a better state than in a worse state for treatment subjects compared to control subjects over the trial follow-up period

Comparison of Win Time Methods and Win Ratio

- Win Time Methods effectively weight death higher than Win Ratio so that overall benefit is highly likely if test is positive for benefit
- Win Time Methods all have higher power than Win Ratio when there is treatment benefit on death
- EWTP weights death highest of these methods and has highest power if treatment benefit on death is substantial
- EWTPR is less biased and can be unbiased with time restriction
- EWTP & EWTPR allow prognostic covariates which can increase power further

Analysis of HF-ACTION

- Three level hierarchy:
  $AE < hospitalization < death$
- Five level hierarchy:
  $AE < 1\ hosp. < 2\ hosp. < 3+\ hosp. < death$

AE is any of 5 pre-specified major adverse events

Analysis of HF-ACTION

Table: Re-analysis of HF-ACTION trial*

Three level hierarchy (AE < hospitalization < death)

	HR†	*EWTP* ≀	*EWTPR*‡
Effect Estimate	0.910	34.2 days	43.7 days
P-value	0.072	0.053	0.052

Five level hierarchy (AE < 1 hosp. < 2 hosp. < 3+ hosp. < death)

	HR†	*EWTP* ≀	*EWTPR*‡
Effect Estimate	0.910	34.8 days	44.0 days
P-value	0.072	0.062	0.065

* All models include pre-specified ischemic heart failure covariate in final models

† hazard ratio for the composite (non-hierarchical) analysis

≀ EWTP using a Markov model

‡ EWTPR using a Markov model, testing with 1000 imputations

Conclusions

- Win Time Methods can greatly reduce the possibility that trials find significant treatment benefit without overall treatment benefit
- This makes trial conclusions much stronger and potentially have greater impact
- Win Time Methods have highest power if treatment has substantial benefit on death, and higher power than pairwise methods (like Win Ratio) when treatment has benefit on death
- Win Time Methods are especially suited for longer followup trials (multiple years), but are appropriate for any trial with hierarchical events
- EWTPR with time restriction can give unbiased estimation of natural estimand (expected excess time in better state over fixed time interval)

References

Troendle JF. et al. (2024) Stat Med. 43: 1920-1932.
O'Connor CM. et al. (2009) J Amer Med Assoc. 301:1439-1450.
Sorajja P et al. (2023) N Engl J Med. 388: 1833-1842.
Finkelstein DM. and Schoenfeld DA. (1999) Stat Med. 18:1341-1354.
Pocock SJ. et al. (2012) Eur Heart J. 33:176-182.
Mao L. (2023) Biometrics. 79: 61-72.

R package **wintime 0.4.4** (2024) J. Troendle and S. Lawrence

Collaborators: Eric Leifer, Song Yang, Dong-Yun Kim, Jungnam Joo, Emma Davies Smith

Recent Advancements in Desirability of Outcome Ranking (DOOR) Methodology

Guoqing Diao

Department of Biostatistics and Bioinformatics
Milken Institute School of Public Health
The George Washington University
Email: gdiao@gwu.edu

SCT 47th Annual Meeting
Phoenix, Arizona, USA

May 19, 2026

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Milken Institute School of Public Health
THE GEORGE WASHINGTON UNIVERSITY

Disclosures

No Relevant Disclosures

Desirability of Outcome Ranking (DOOR)

STATISTICS IN BIOPHARMACEUTICAL RESEARCH
2016, VOL. 8, NO. 4, 386–393
http://dx.doi.org/10.1080/19466315.2016.1207561

Taylor & Francis
Taylor & Francis Group

Using Outcomes to Analyze Patients Rather than Patients to Analyze Outcomes: A Step Toward Pragmatism in Benefit:Risk Evaluation

Scott R. Evans[a,b] and Dean Follmann[c]

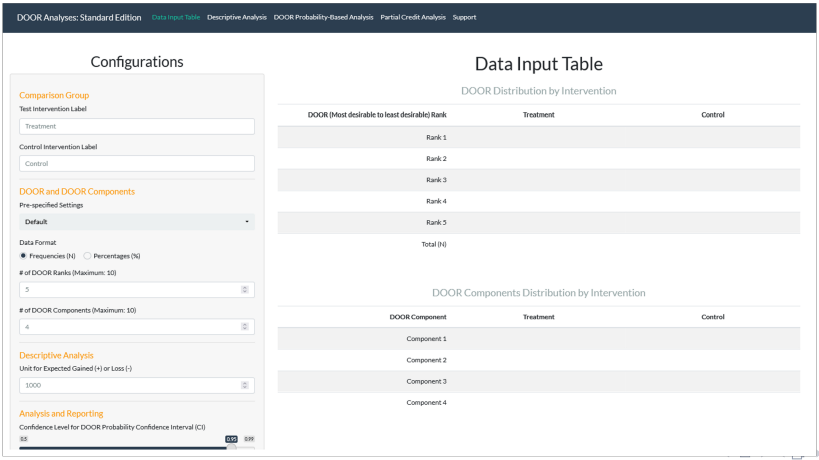
[a]Department of Biostatistics, Harvard University, Boston, MA, USA; [b]Center for Biostatistics in AIDS Research, Harvard University, Boston, MA, USA; [c]National Institute of Allergy and Infectious Diseases (NIAID) of the National Institutes of Health (NIH), Bethesda, MD, USA.

- The desirability of outcome ranking (DOOR) is a patient-centric paradigm for the design, data monitoring, analysis, interpretation, and reporting of clinical trials and other research studies based on benefit:risk evaluation
- Patient-centric in this context means that the DOOR paradigm uses outcomes to analyze patients rather than patients to analyze outcomes (Evans et al., 2015; Hamasaki et al., 2025). DOOR Shiny app and R package are available

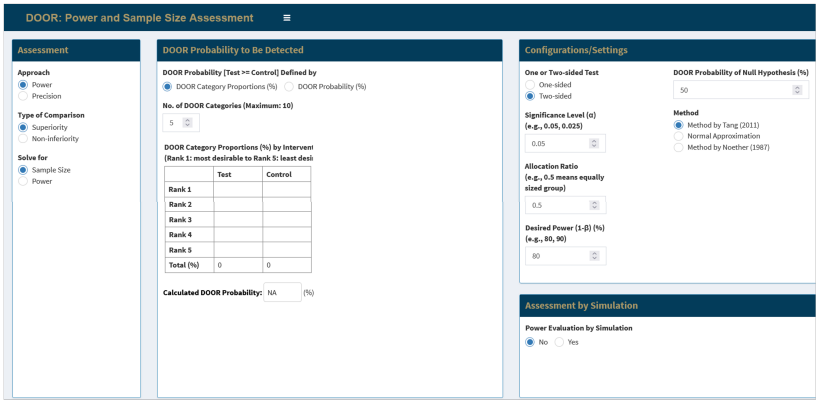
DOOR Analysis

- Guiding principles for replicability, pragmatism, and robustness
- Two complimentary analyses
  1. Rank-based
     - Estimate the **DOOR probability** $D = Pr(Y_t < Y_c) + \frac{1}{2} Pr(Y_t = Y_c)$: the probability that a patient from treatment has a more desirable outcome than a patient on control (equivalent distributions imply 50%). A smaller rank value (e.g., Rank 1) indicates a more desirable outcome
  2. Partial credit (grade-based analyses)
- Analysis of component outcomes is an integrated part of the evaluation

Online tools for implementing DOOR analyses



Online tools for implementing DOOR analyses



Related Work

- Win Ratio (Pocock et al., 2012; Oakes, 2016; Bebu & Lachin, 2016; Mao & Wang, 2020; Mao, 2021; Haque et al. 2026; etc.)

$$Pr(Y_t < Y_c) / Pr(Y_t > Y_c)$$

- Generalized Pairwise Comparisons (GPC) (Buyse, 2010; Péron et al., 2018; Giai et al., 2021; Verbeeck et al., 2023; etc.)

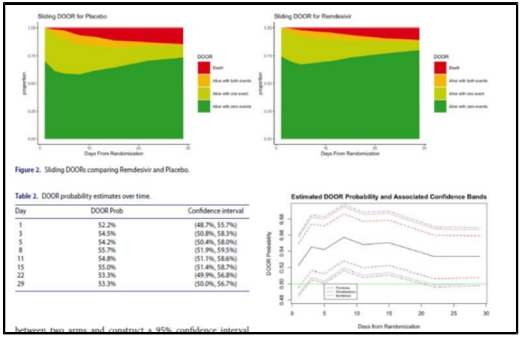
$$\text{Net Benefit} = Pr(Y_t < Y_c) - Pr(Y_t > Y_c)$$

- Win Odds (Brunner et al., 2021)

$$\left\{ Pr(Y_t < Y_c) + \frac{1}{2} Pr(Y_t = Y_c) \right\} / \left\{ Pr(Y_t > Y_c) + \frac{1}{2} Pr(Y_t = Y_c) \right\}$$

- Win Time (Troendle et al., 2024)
- Reviews and comparisons of various methods (Dong et al., 2023; Barnhart et al., 2025; Backer et al., 2026)

Sliding DOOR



Sliding DOOR

- View DOOR outcome as a longitudinal patient state
- Methods for:
  - Repeated measures
    - Simultaneous confidence bands
  - Monotone progression
    - Multidimensional RMST
  - Non-monotone levels
    - Anthology of Patient Stories

Cardiovascular Prevention DOOR (CAR-DOOR?)

- Events of interest
  - Death
  - Stroke
  - MI
  - Major bleeding
- DOOR
  - Alive with no events
  - Alive with 1 non-disabling event
  - Alive with >1 non-disabling event
  - Alive with disabling event
  - Death

Article

CLINICAL TRIALS

Clinical Trials
2026, Vol. 23(1) 23–32

DOI: 10.1177/17407745251385582
journals.sagepub.com/home/ctj
Sage

Desirability of outcome ranking (DOOR) analysis for multivariate survival outcomes with application to ACTT-1 trial

Shiyu Shu, Guoqing Diao, Toshimitsu Hamasaki and Scott Evans

Abstract

Background: Desirability of outcome ranking (DOOR) is a paradigm for the design, monitoring, analysis, interpretation, and reporting of clinical trials based on patient-centric benefit-risk evaluation, developed to address limitations of existing approaches and advance clinical trial science. The first step in implementing DOOR is defining an ordinal DOOR outcome representing a global patient-centric response, a cumulative summary of the benefits and harms for an individual patient. This article aims to develop an analysis methodology for the setting where the DOOR outcome is a progressive time-varying state, and there is interest in event times and times that patients spend in more and less desirable states.

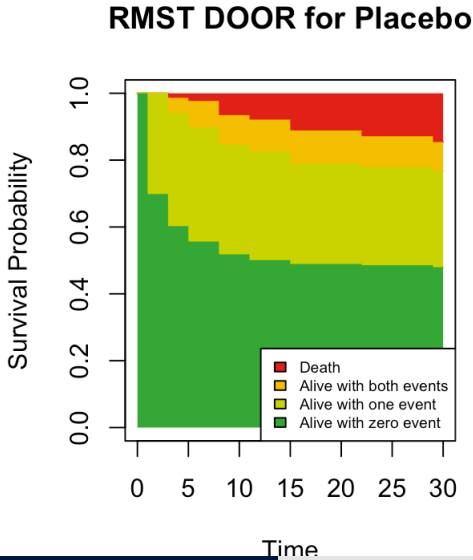
Methods: We develop methods to estimate and make inferences about the temporal treatment effects. If the k-levels of the DOOR outcome are monotone, then k – 1 non-overlapping Kaplan-Meier survival curves can be estimated and plotted. The areas under the curves asymptotically follow a multivariate Gaussian distribution. We apply restricted mean

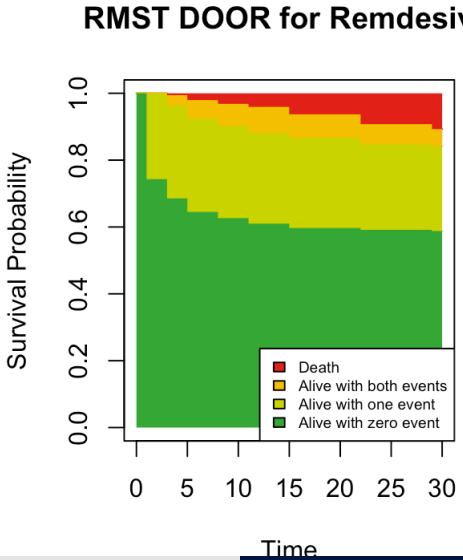
Connection between multivariate survival endpoints and DOOR endpoints

- Consider qualifying events: 1. Stroke 2. Severe Bleeding 3. Disability 4. Death

1. Survived with 0 events
2. Survived with 1 event w/o disability
3. Survived with >1 event w/o disability
4. Survived with overall disability
5. Death

1. T_1 = event-free survival time, or time to the first occurrence of any qualifying events
2. T_2 = time to both stroke and severe bleeding, or disability, or death
3. T_3= time to disability
4. T_4 = time to death


RMST DOOR for Placebo
Survival Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30
Time
Death
Alive with both events
Alive with one event
Alive with zero event


RMST DOOR for Remdesivir
Survival Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30
Time
Death
Alive with both events
Alive with one event
Alive with zero event

On Cluster Randomized Trials with the Desirability of Outcome Ranking (DOOR) Endpoints

Wanying Shao[1], Toshimitsu Hamasaki[1], Scott Evans[1] & Guoqing Diao[1*]

[1] Department of Biostatistics and Bioinformatics

The George Washington University, Washington, D.C.

Abstract

Cluster randomized trials are widely used when individual randomization is logistically infeasible or when correlations between observations cannot be ignored, especially in fields such as ophthalmology, infectious disease, vaccine research, and social

Cluster randomized trials (CRT)

- Cluster randomized trials are widely used in ophthalmology, infectious diseases, vaccine research, and sociological studies
- In CRTs, observations within the same cluster are more alike than observations across clusters. Failing to account for the intracluster correlation coefficient (ICC) will underestimate the variance of the parameter estimator for the intervention effect and lead to misleading results (Jung, 2024)

Motivations

- Parallel cluster randomized trials
- Additional considerations and designs in ophthalmology studies: each eye of a patient may receive different treatments, which requires attention to the correlation between the eyes (Rosner, 1982). In the study described by Pall et al. (2019):
  1. Contralateral/split-eye group, in which subjects received the test lens in one eye and the control lens in the contralateral eye;
  2. Test lenses group in which both eyes received the same intervention;
  3. Control lenses group, in which both eyes received the control lens.
- Milking In Non-vigorous Infants (MINVI) study: a cluster randomized crossover trial
- Most recent work on win statistics in CRTs focuses on parallel CRTs (e.g., Fang et al., 2025)

What's New?

- A suite of new methods to extend DOOR to cluster randomized trials based on properties of U-statistics and influence functions to estimate within-cluster and between-cluster treatment effects
- A unified framework that can be applied in:
  - One-group randomization (parallel cluster randomized trials)
  - Two-group randomization (crossover trials, or contralateral/split-eye designs)
  - Mixture randomization (50% one-group randomization, and 50% two-group randomization)
  - Small number of clusters
  - Small cluster sizes

Notation

Suppose that there are n clusters with m_i subjects in the ith cluster. The DOOR endpoint has K levels.

- A_{ij}: Treatment group indicator of the jth subject in the ith cluster
- Y_{ij}: DOOR rank of the jth subject in the ith cluster
- D_{wi}: DOOR probability for cluster i
- D_w: Within-cluster DOOR probability
- D_b: Between-cluster DOOR probability

DOOR probability

For the ith cluster, the DOOR probability is defined as

$$D_{wi} = E\{I(Y_{ij} < Y_{ij'}) + I(Y_{ij} = Y_{ij'})/2\},$$

for a randomly selected pair such that $A_{ij} = 1$ and $A_{ij'} = 0$.
We estimate D_{wi} by,

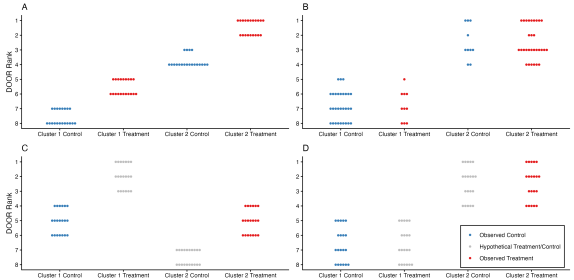
$$\widehat{D}_{wi} = \frac{1}{m_{i1} m_{i2}} \sum_{j=1}^{m_i} \sum_{j'=1}^{m_i} \phi(Y_{ij}, Y_{ij'}),$$

where $\phi(Y_{ij}, Y_{ij'}) = A_{ij}(1 - A_{ij'})\{I(Y_{ij} < Y_{ij'}) + I(Y_{ij} = Y_{ij'})/2\}$, and $m_{i1} = \sum_{j=1}^{m_i} A_{ij}$ and $m_{i2} = m_i - m_{i1}$ are the numbers of subjects in the treatment group and control group, respectively.

- **Within-Cluster DOOR Probability**
  - Inverse variance weighted within-cluster DOOR probability estimator $\widehat{D}_w$
  - Sample size weighted within-cluster DOOR probability estimator $\widetilde{D}_w$
  - Small-sample correction
- **Between-Cluster DOOR Probability**
  - $$D_b = E\{I(Y_{ij} < Y_{i'j'}) + I(Y_{ij} = Y_{i'j'})/2\},$$

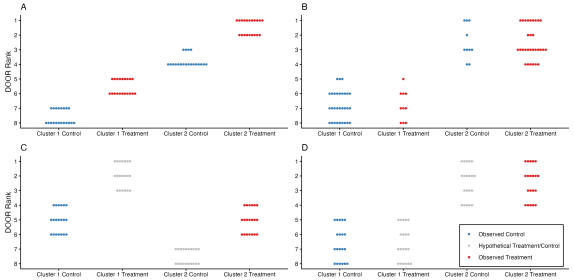
    for a randomly selected pair such that $A_{ij} = 1$, $A_{i'j'} = 0$, and $i \neq i'$
  - $\widehat{D}_b$: Mann-Whitney statistic for comparisons between clusters (Larocque et al., 2010)

Within-Cluster versus Between-Cluster DOOR Probabilities



- In panel A, a clear treatment effect within each cluster. D_b fails to detect, leads to a false negative result
- In panel B, no treatment effect within each cluster. D_b leads to a false positive result

Within-Cluster versus Between-Cluster DOOR Probabilities



- Gray dots indicate the potential outcomes of patients
- In panel C, D_b leads to a false negative result. In panel D, D_b leads to a false positive result
- Highlight that the between-cluster DOOR probability can be driven by cluster-level factors that are unrelated to the direct effect of treatment on an individual

- **Hypothesis Testing**
  - Assess the between-cluster variability, $H_0 : D_b - D_w = 0$, $\widehat{W}_v$
  - Simultaneous testing using L_∞ norm, $H_0 : D_b = D_w = 0.5$, $\widehat{W}_{max} = \max(\widehat{W}_w, \widehat{W}_b)$
  - Weighted average of D_b and D_w $H_0 : D_b = D_w = 0.5$, $\widehat{W}_{wt}$

Guidance for the choice of the proposed methods in different practical settings.

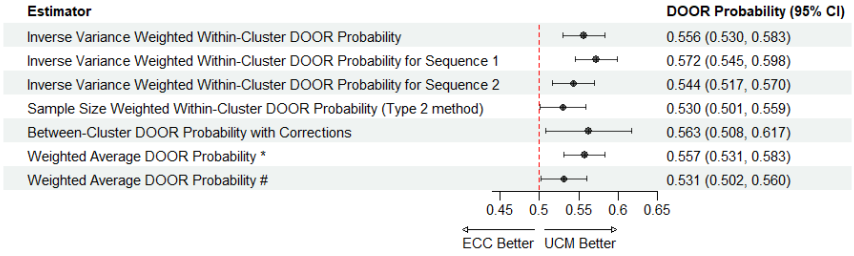
n	m_i	**Within-cluster estimator**	**Variance Estimator**
Small	Large and small within-cluster correlation	$\widehat{D}_w$	Estimated by $1/\sum_{i=1}^{n} m_i \widehat{\sigma}_i^{-2}$ Same as the Fixed-Effects Meta-Analysis approach
Small	Large	$\widetilde{D}_w$	[Type 2 method] Estimated by $\sum_{i=1}^{n} \tilde{w}_i^2 \sigma_i^2$
Small	Moderate/Large	$\widetilde{D}_w$	[Type 3 method] Use influence functions with small sample size corrections
Large	Small/Large	$\widetilde{D}_w$	[Type 1 method] Use influence functions, can estimated using the empirical version $\sum_{i=1}^{n} \tilde{w}_i^2 (\widehat{D}_{wi} - \widetilde{D}_w)^2$

Guidance

n	m_i	**Between-cluster estimator**	**Variance Estimator**
Small	Moderate/Large	$\widehat{D}_b$	Use influence functions with small sample size corrections
Large ($n \geq 15$)	Small/Large	$\widehat{D}_b$	Use influence function

Application

- The MINVI study is a cluster randomized crossover trial conducted among 1730 newborns from 10 medical centers between January 2019 and May 2021 (Katheria et al., 2023). 1524 newborns with complete measurements on all DOOR components
- The study hypothesized that umbilical cord milking (UCM) would reduce admission to the neonatal intensive care unit (NICU) compared with early cord clamping (ECC)
- Hospitals were randomized in a 1:1 ratio to UCM or ECC in period one, then crossed over to the other intervention in period two.



Discussion

- We have developed methodologies for cluster randomized trials with the DOOR endpoints, taking into account different designs, and we have introduced two estimands: the within-cluster and the between-cluster DOOR probabilities, and we have developed estimation and inference procedures for both estimands
- Limitations:
  - Focus on analyzing data, haven't developed methods for sample size and power calculations
  - Assume equal weights for each comparison between two clusters when estimating D_b
  - Assume that there are no temporal effects under the crossover design in the real application

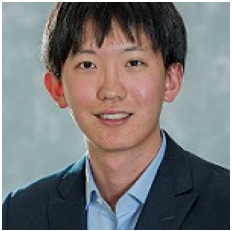
Ongoing and Future Research

- Interim monitoring is a future direction, will extend the group-sequential designs for eye trials (Diao et al., 2025) to generic cluster-randomized trials
- An extension of the proposed methods to stepped-wedge cluster randomized trials
- Causal methods for observational studies (Shu et al., 2026; Feng et al., 2026)
- Sequential, Multiple Assignment, Randomized Trial (SMART) design with DOOR endpoints

Key References

- Evans, S.R., Rubin, D., Follmann, D., Pennello, G., Huskins, W.C., Powers, J.H., Schoenfeld, D., Chuang-Stein, C., Cosgrove, S.E., Fowler Jr, V.G. and Lautenbach, E., 2015. Desirability of outcome ranking (DOOR) and response adjusted for duration of antibiotic risk (RADAR). Clinical Infectious Diseases, 61(5), pp.800-806.
- Hamasaki, T., He, Y., Wu, Q., Howard-Anderson, J., Boucher, H.W., Doernberg, S.B., Holland, T.L., Powers III, J.H., Wang, J., Diao, G. and van Duin, D., 2025. A patient-centric paradigm and tool for clinical research: the DOOR is open. Antimicrobial agents and chemotherapy, 70(1), pp.e01478-25.
- Shu, S., Diao, G., Hamasaki, T. and Evans, S., 2025. Longitudinal benefit: risk analysis through the desirability of outcome ranking (DOOR) with application to ACTT-1 trial. Statistics in Biopharmaceutical Research, 17(3), pp.488-495.
- Shao, W., Hamasaki, T., Evans, S. and Diao, G., 2026. On Cluster Randomized Trials with the Desirability of Outcome Ranking (DOOR) Endpoints. arXiv preprint arXiv:2604.24032.

Acknowledgement



Thank You!

Duke Clinical Research Institute

Discussion on Innovative Methods for Composite Endpoint

Huiman Barnhart, PhD
James B. Duke Distinguished Professor
Duke University

Society for Clinical Trials Annual Meeting
May 19, 2026

Presentation by Dr. Rejuan Haque

- Thank you for the gentle and nice Introduction of Win Ratio, its applications, and software
- Mentioned the Extension to Win Odds and Net Benefit
- Comment: May Include $DOOR = P_{win} + 0.5\ P_{tie}$ on hierarchical endpoints (Barnhart et al., 2025)
- Potential other pitfalls of WR and other Win Measures
  - Win is a Win, and it doesn't matter where you win - in the first endpoint, second, or third endpoints?
  - Least important endpoint can still dominate the win - complicating interpretation; how do you label the drug?
  - Usual WR does not always use all events – a subject can lose along the way, but ultimately win at the end

Presentation by Dr. Rejuan Haque

- Potential other pitfalls of WR and other Win Measures
  - Usual WR does not always use all events
    - Methods that use all events at the times of occurring:

      Win Time (Troendle et al. 2024) – summary of all NBs at all time points (more later)

      Weighted Composite Endpoints (Bakal et al, 2014, Nabipoor et al. 2023) – similar to KM curves

      Longitudinal DOOR (Shiyu Shu and Guoqing Diao et al., 2025) – DOORs over time
  - Dependency on follow-up time – studies are not comparable without the same FU time
    - Fix: assuming that the win measure is the same regardless of FU time. In practice, this is usually not true.
  - Dependency on censoring distribution – an issue with an estimand
    - Time restricted WR (Wang et al., 2025)
    - Need estimation method to deal with missing data in addition to censoring

Presentation by Dr. Rejuan Haque

- With increased use of WR in clinical research program, study design issues are important
  - Trial Design with Sample Size Calculations
  - Covariate Adjustment
    - Win Fraction Regression, requiring WR not dependent on FU time
  - Study Monitoring – WR with interim data is not the same WR at the end of study as WR often depends on FU time

Presentation by Dr. Rejuan Haque

- Trial Design with Sample Size Calculations
  - Simulation-based approach

    Unclear Assumptions – need to specify multivariate distribution – Oftentimes, the clinically significant WR used in the alternative hypothesis is unknown not reported.

    Time Consuming - Iterative process to obtain the sample size to achieve the pre-specified power (Redfors et al. 2020)

Presentation by Dr. Rejuan Haque

- Trial Design with Sample Size Calculations
  - Formula-based approach

    Yu and Ganju (2022, Statistics in Medicine)

    - Simple formula and quick and easy to compute,
    - Hard to specify the parameters of clinically significant WR, probability of ties.
    - Can be conservative due to variance approximation

Presentation by Dr. Rejuan Haque

- Trial Design with Sample Size Calculations
  - Formula-based approach

    Barnhart et al. (2025, Statistics in Medicine)

    - Use Yu and Ganju's formula for quick and easy computation
    - A simple and meaningful way to specify clinically significant WR and probability of tie

      Linking marginal parameters for each endpoint, e.g., HR, OR, Mean (SD), etc. with the WR and probability of ties under independence assumption

    - Available R-shiny app and R package for easy computation
    - Serve as good starting sample size for simulation-based approach, if not comfortable with independence assumption

Presentation by Dr. James Troendle

- Win Time method is an effective way to utilize all events at the time occurrences of these events, while preserving the hierarchy at each time point
- Win Time fixes one of the potential pitfalls by using all EVENTS
- Win/Lose/Tie is considered multiple times at each time point unit until censoring or planned FU time
  - PWT – pairwise win time, counting approach
  - EWT – expected win time, probability approach against Control Population
  - EWTP – expected win time against the Trial Population, regression approach
  - EWTPR – expected win time against trial population with redistribution (with multiple imputations to deal with censoring issue)
- If there is no censoring and everyone is followed to a fixed FU time
  - PWT=EWT=RMT-IF (=EWTP=EWTPR?)

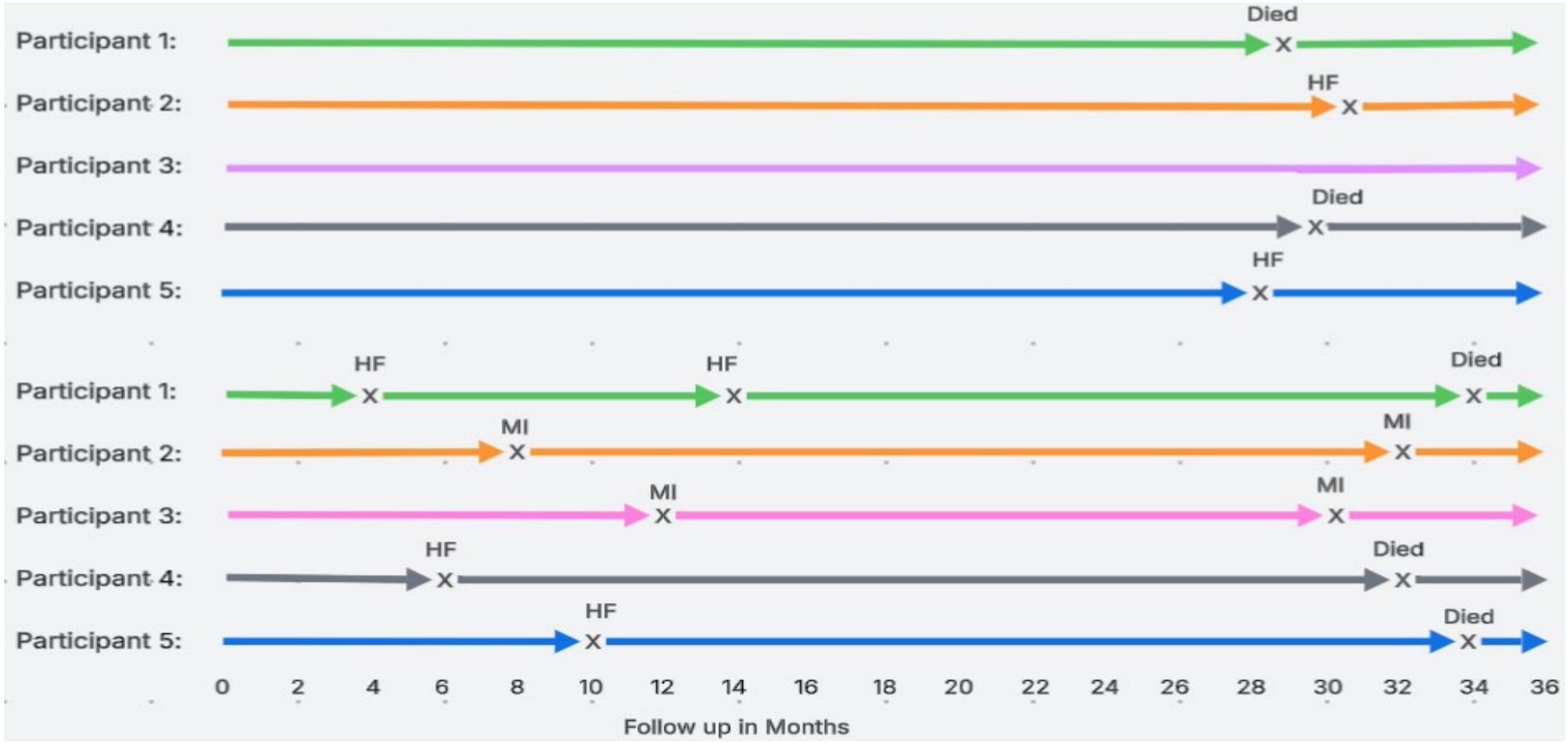
  with RMT-IF as the restricted mean time in favor of treatment (Mao, 2023)

My Take on Win Time with a Toy Example:

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

Active Arm

Control Arm



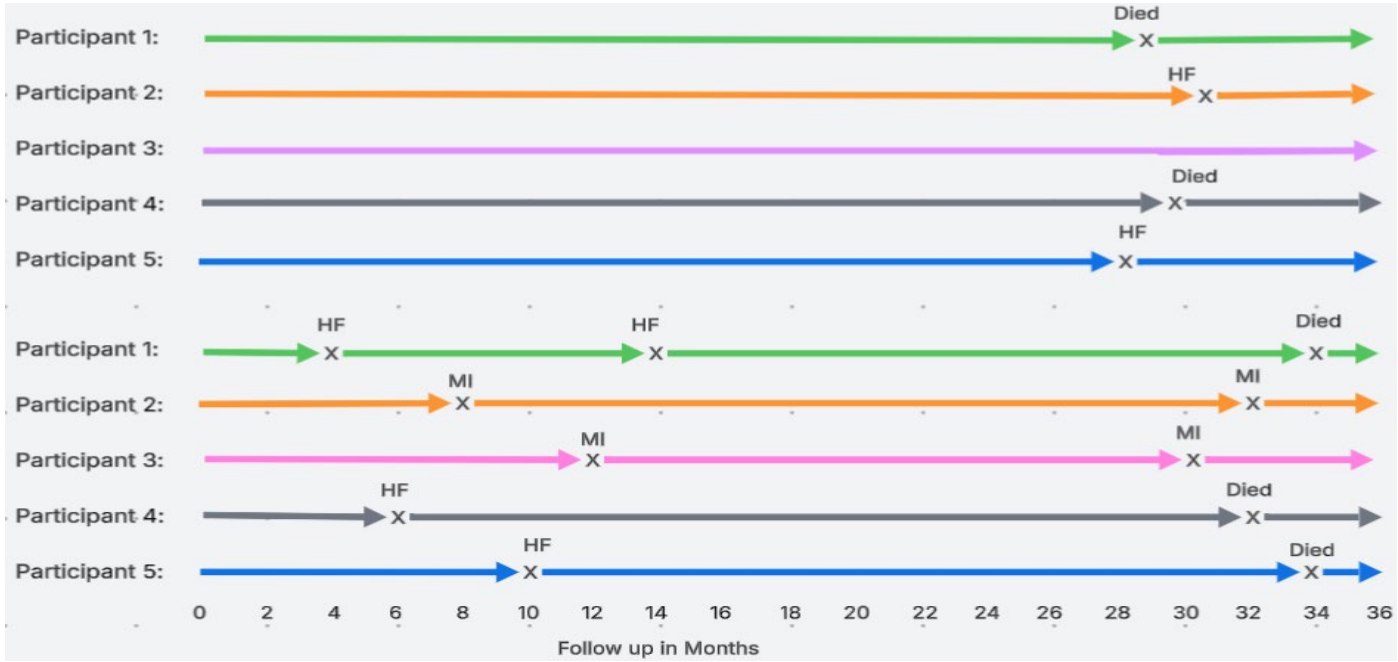
Hierarchy: Death > 2HF > 1HF > 2MI > 1MI

Toy Example:

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

Active Arm

Control Arm

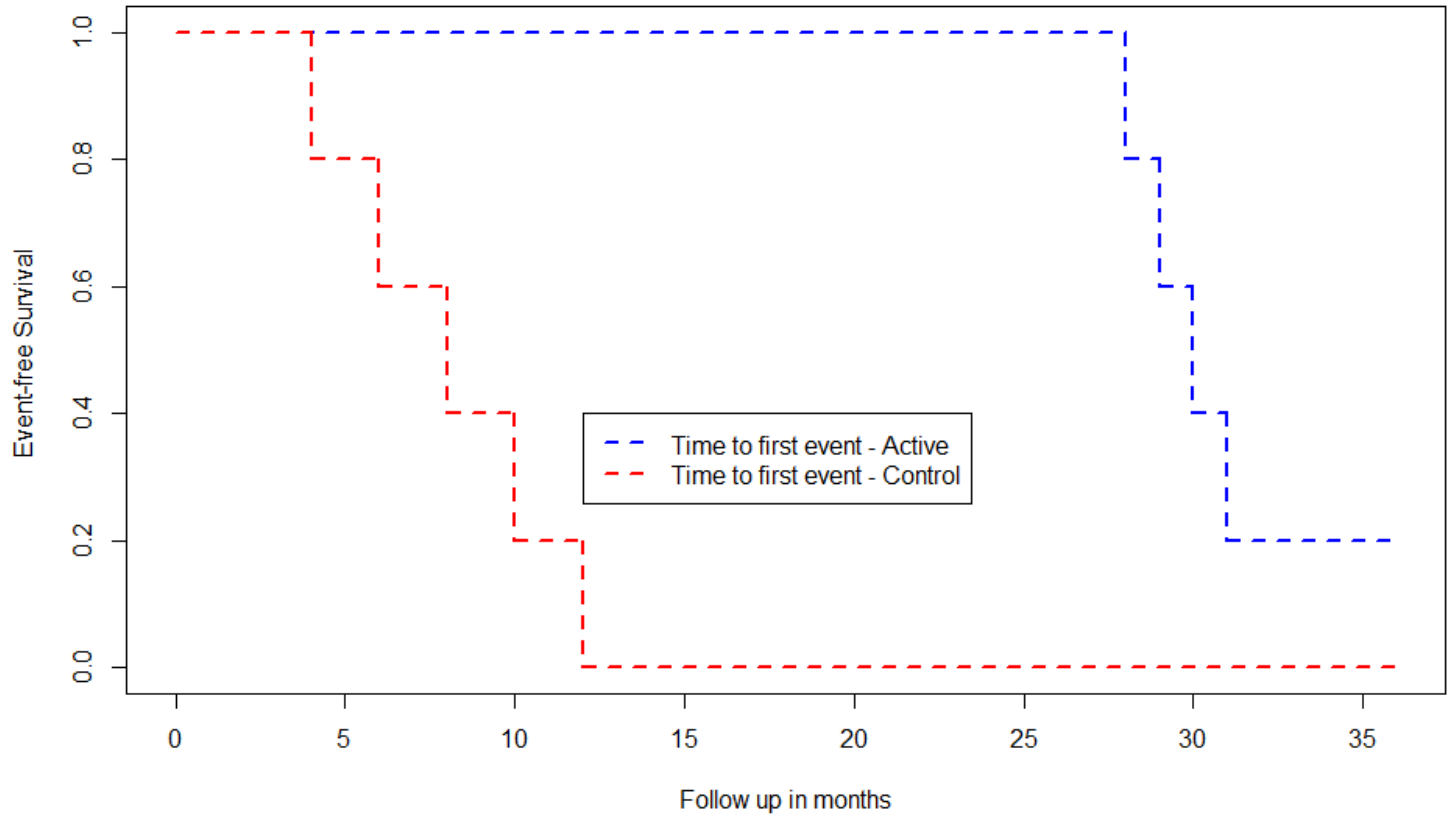


Paired subjects (1,1) over 36 months

- In WR or NB, one win/lose/tie over 36 months, active subject loses and control subject wins
- In win time, multiple win/lose/tie at every month until month 36, active subject wins in the first 28 of the 36 months, then loses in the last 8 months – a net win of 20 months

Toy Example:

Time to first event:



Hierarchy: Death > HF > MI.

WR=11/14=0.79

Hierarchy: Death > 2HF > 1HF > 2MI > 1MI, 36-month FU

PWT=EWT=RMT-IF=20.52 months

Presentation by Dr. James Troendle

- Question 1: What are the Estimands as population parameters as the concepts are defined as estimators? How do you decide which one to use in practice?
  - PWT – pairwise win time, depends on individual observation time
  - EWT – expected win time, against Control Population
  - EWTP – expected win time against Trial Population
  - EWTPR – expected win time against Trial Population with redistribution (deal with censoring issue)
- The state probabilities in the trial population seems to be the weighted average of the state probabilities in the active and control groups.
  - It is not clear if EWT=EWTP without censoring?

  EWT is based on state probabilities against Control Population

  EWTP is based on linear regression against Trial Population (treatment dependent?)

Presentation by Dr. Guoqing Diao

- Focused on DOOR (Desirability of Outcome Ranking) measure
- Nice presentation on using with-cluster and between cluster DOORs to assess treatment effect within cluster and between clusters, respectively
- Should we expect to have the same treatment effect for those within cluster and between cluster?
  - Testing $H_0: \mathrm{DOOR_w} = \mathrm{DOOR_b}$, it may be due to low power. Should you always be polling $\mathrm{DOOR_w}$ and $\mathrm{DOOR_b}$ together?
  - For patients between clusters, they may experience different clinical care environment. Different characteristics between clusters may contribute to the estimation of $\mathrm{DOOR_b}$

Presentation by Dr. Guoqing Diao

- For randomization at the cluster lever, everyone in the same cluster receives the same treatment. How about considering a different $\mathrm{DOOR_b}$ at cluster level, rather than at patient level, e.g.,

$$D_{b,clus} = E_{ii'}E_{jj'}\left\{I(Y_{ij} < Y_{i'j'}) + \frac{I(Y_{ij} = Y_{i'j'})}{2}\right\},$$

$$\widehat{D}_{b,cluster} = \frac{\sum_{i=1}^{n}\sum_{i'=1}^{n} A_i(1-A_{i'})\frac{1}{m_i m_{i'}}\sum_{j=1}^{m_i}\sum_{j'=1}^{m_{i'}}\phi(Y_{ij}, Y_{i'j'})}{\sum_{i=1}^{n}\sum_{i'=1}^{n} A_i(1-A_{i'})}$$

Presentation by Dr. Guoqing Diao

- DOOR is usually based on one ordinal endpoint with small number of categories, typically by combining different categories among combining efficacy and safety endpoints
  - Small number of categories → easier to interpret
  - Combining multiple categories may dilute the treatment effect?

  For example, 3 efficacy endpoints (Yes/No) and 2 safety endpoints (Yes/No)

  Combinaing $2^5 = 32$ categories → 6 categories

  may miss the treatment effect in some combinations of efficacy and safety categories