Duke Clinical Research Institute

Trial Design with Win Ratio or Win Odds Based on Hierarchical Endpoints

Huiman Barnhart
Joint work with
Yuliya Lokhnygina, Roland Matsouaka, and Frank Rockhold

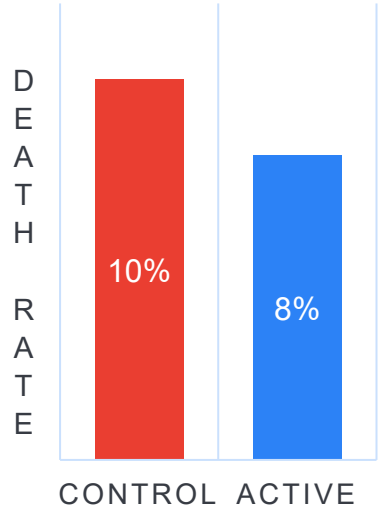
SCT 46th Annual Meeting
May 21, 2025

A Typical Scenario in Designing a Trial with Hierarchical Endpoints

1. Time to death by 1 year

Control (10%) vs. Active (8%):

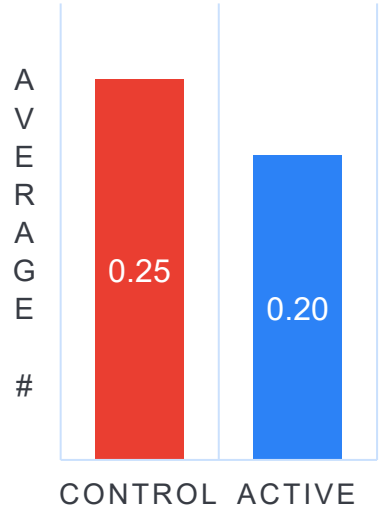
20% reduction or HR=0.79



2. Average Number of hospitalizations by 1 Yr.

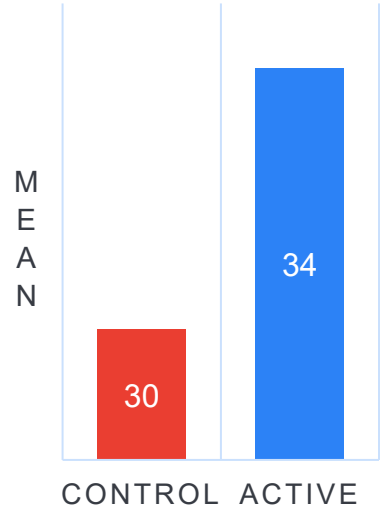
Control (0.25) vs. Active (0.2)

20% reduction or Rate Ratio=0.8



3. Mean (SD) change of KCCQ score at 1 year (threshold > 5 to win)

Control 30 (24) vs. Active 34 (24)



What is the overall Win Ratio? How to justify it?

Win Ratio (WR)

- Introduced in Pocock et al. (2012)
- 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."

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

Null Hypothesis: WR=1

- WR depends on time horizon (S) where treatments are compared

Typical Approach in determining sample size via simulations

- Need to specify a multivariate distribution of the hierarchical endpoints for data generation.
  - Difficult to do and it needs reasonable assumption on correlations
- What sample size should be used to simulate data? An iterative process.
- Time-consuming due to a large number of pair comparisons, e.g., with 5,000 simulations, it may take several days to run for one scenario with a sample size of 3000).
- What's the expected overall win ratio? Is it justifiable?
- What's the expected probability of ties? Is it justifiable?

Determining Trial Size with Formula

- The following sample size and power formula (Yu and Ganju, 2022) is recently available for WR in PASS 2024

$$N_{WR} = \frac{4(1+p_{tie})\left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right)^2}{3r(1-r)(1-p_{tie})\left(\log\left(WR(S;Y_1,\ldots,Y_K)\right)\right)^2}, \qquad (1)$$

$$\text{Power}_{\text{WR}} = 1 - \Phi\left(Z_{1-\frac{\alpha}{2}} - abs(\log(WR(S;Y_1,\ldots,Y_K)))\right)\sqrt{\frac{3r(1-r)N(1-p_{tie})}{4(1+p_{tie})}}). \qquad (2)$$

- **Need to know the justifiable effect size (the overall WR) and the probability of ties**
- **A typical convenient fix: I will add a continuous endpoint at the end of hierarchy so that the probability of ties is zero. It may not always be a good solution**

Our New Tool

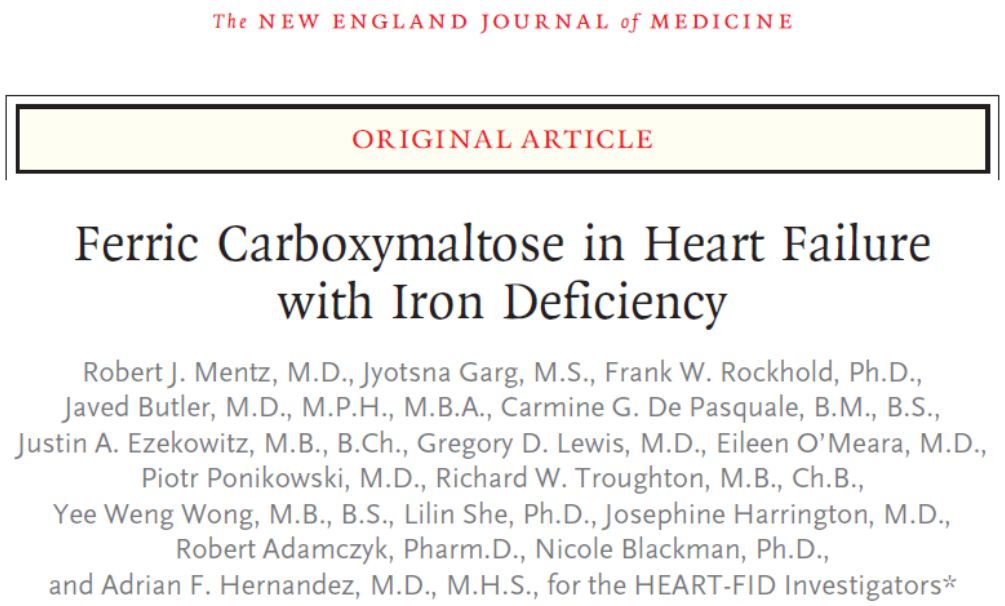
- Assume **INDEPENDENCE** of endpoints
  - WR is a weighted average of marginal WRs with weights depending on probability of loss and probability of ties of proceeding endpoint.

$$\boldsymbol{WR}(S;\ Y_1, \dots, Y_K) = \boldsymbol{w_{Y_1}} * \boldsymbol{WR_{Y_1}} + \boldsymbol{w_{Y_2}} * \boldsymbol{WR_{Y_2}} + \cdots + \boldsymbol{w_{Y_K}} * \boldsymbol{WR_{Y_K}}$$

- Benefits of the formula
  - A way to link $\boldsymbol{WR}(S;\ Y_1, \dots, Y_K)$ via marginal $\boldsymbol{WR}$s as marginal $\boldsymbol{WR}$ corresponds to traditional effect size that is easy to understand and specify
  - A way to justify the magnitude of win ratio as clinically significant
- Properties of the formula
  - $\boldsymbol{WR}(S;\ Y_1, \dots, Y_K)$ is between the minimal and maximum of the marginal WRs
  - Adding a less effective endpoint would decrease an overall win ratio and vice versa
- Probability of ties $\boldsymbol{p_{tie}} = \prod_{k=1}^{K} \boldsymbol{P}(\textbf{ties in } \boldsymbol{Y_k})$

Paper published in Aug. 2023 by NEJM.

DCRI HEART-FID Trial



The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

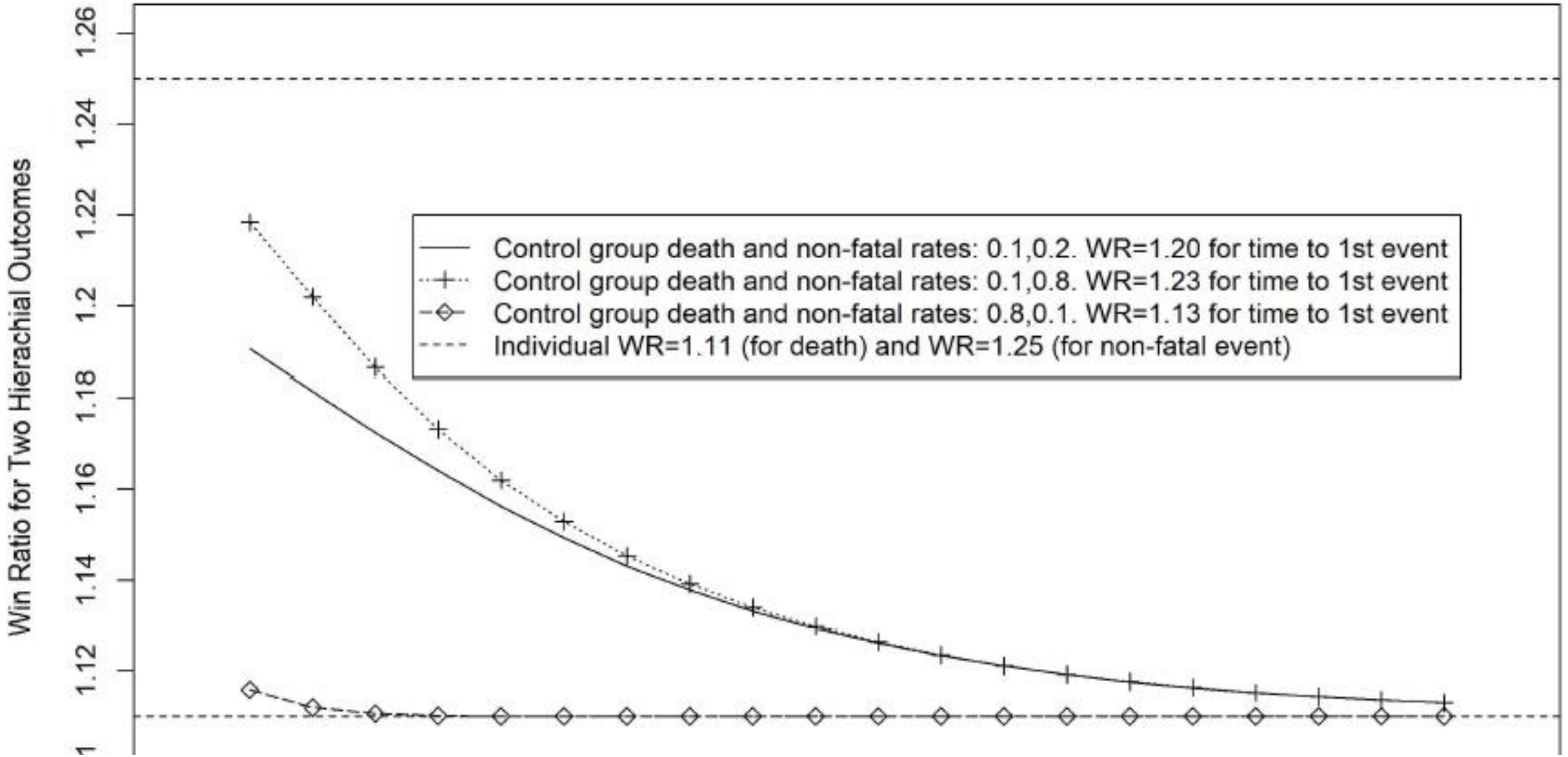
Ferric Carboxymaltose in Heart Failure with Iron Deficiency

Robert J. Mentz, M.D., Jyotsna Garg, M.S., Frank W. Rockhold, Ph.D., Javed Butler, M.D., M.P.H., M.B.A., Carmine G. De Pasquale, B.M., B.S., Justin A. Ezekowitz, M.B., B.Ch., Gregory D. Lewis, M.D., Eileen O'Meara, M.D., Piotr Ponikowski, M.D., Richard W. Troughton, M.B., Ch.B., Yee Weng Wong, M.B., B.S., Lilin She, Ph.D., Josephine Harrington, M.D., Robert Adamczyk, Pharm.D., Nicole Blackman, Ph.D., and Adrian F. Hernandez, M.D., M.H.S., for the HEART-FID Investigators*

Hierarchical Endpoints

1. Time to death by 1 year
2. Number of heart failure hospitalizations by 1 year
3. Change of 6-minute walk distance from baseline to 6-months

Findings with our Tool: WR depends on Follow Up Time
Two Time to Event Endpoints



Heart-FID Data for each of the three endpoints

Observed Death Rate at 1 Year		Observed Averaged Number (Adjusted for FU Time) of Hospitalizations for Heart Failures at 1 Year		Mean (SD) of Observed and Imputed Change From Baseline to Month 6 in 6-min Walk Distance	
FCM	Placebo	FCM	Placebo	FCM	Placebo
0.086	0.103	0.257	0.332	-22.22 (106.83)	-24.02 (101.17)

Design a New Trial with Heart-FID Data as Preliminary Data

Hierarchical Endpoints	Observed Marginal Win Ratio	Calculated Marginal Win Ratio	Calculated Weights	Observed Overall Win Ratio	Calculated Overall Win Ratio
(1) Time to death by 1 Year	1.203	1.209	0.175	1.106	1.149
(2) No. of Hospitalizations for Heart Failure by 1 Year	1.137	1.343	0.298		
(3) Change from Baseline to Month 6 in 6-Min. Walk Distance	1.096	1.02	0.526		

Design a New Trial with Heart-FID Data as Preliminary Data

Total Sample Size	Correlations b/w Three Endpoints	Average Simulated Win Ratio	Calculated Power	Simulation- based Power**
3064	0.0, 0.0, 0.0	1.150	0.919	0.918
3064	-0.13, 0.50, -0.11*	1.151	0.919	0.919
2488	0.0, 0.0, 0.0	1.151	0.85	0.856
2488	-0.13, 0.50, -0.11*	1.151	0.85	0.857
2176	0.0, 0.0, 0.0	1.150	0.80	0.806
2176	-0.13, 0.50, -0.11*	1.151	0.80	0.807

**Observed correlations*
*** Based on 20,000 simulated data sets*

Example 1: Two Endpoints vs. Three Endpoints?

Hierarchical Endpoints	Calculated Marginal Win Ratio
(1) Time to death by 1 year	1.209
(2) No. of Hospitalizations for Heart Failure	1.343
(3) Change from Baseline to Month 6 in 6-Min Walk Distance	1.02

	Calculated Win Ratio	Probability of Ties	Power with N=3064
With two endpoints	1.293	0.495	0.948
With Three Endpoints	1.149	0.00	0.916

Higher power with 2 endpoints than with 3 endpoints

Investigation on Independence Assumption

- With positive correlation, if the pair of subjects are tied on the first outcome, they are also likely to be tied on the second outcome and thus

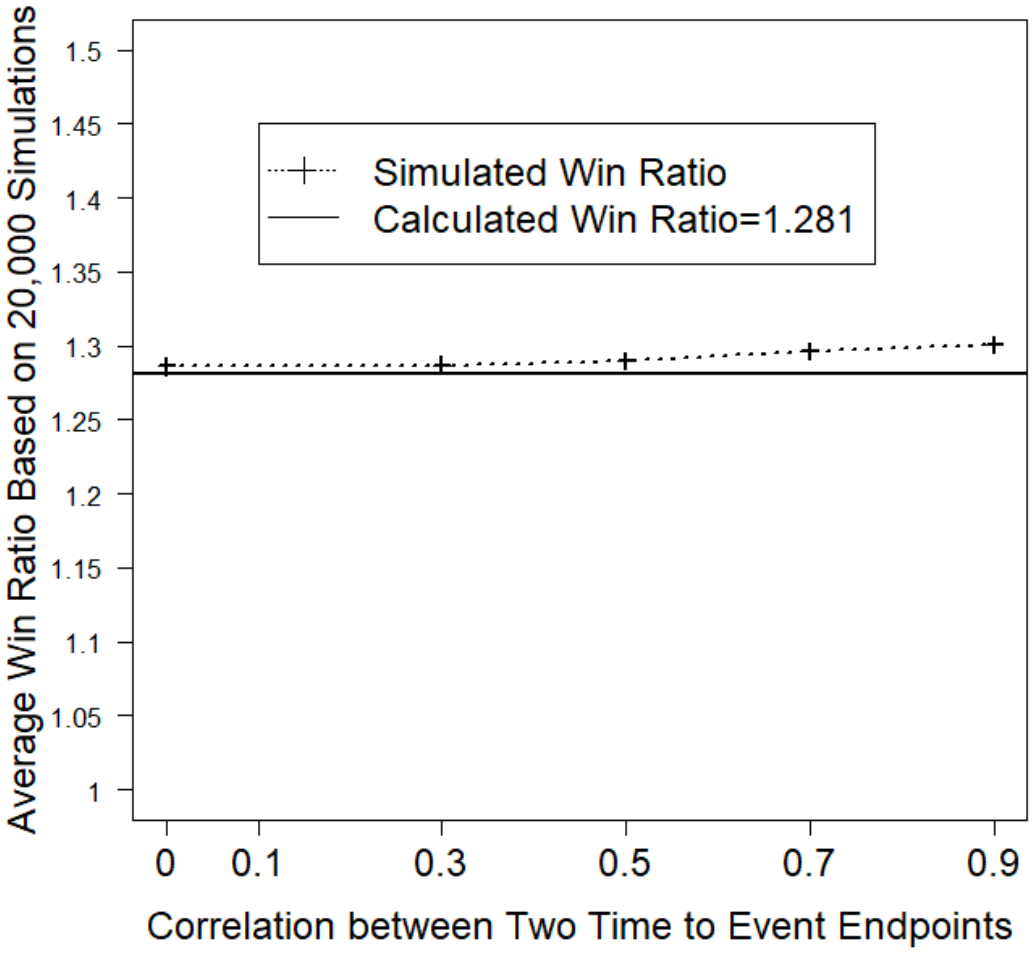
$$Prob(A\ wins|FU = S) \leq Prob(A\ wins|FU = S, independence)$$
$$Prob\ (B\ wins|FU = S) \leq Prob(B\ wins|FU = S, independence)$$

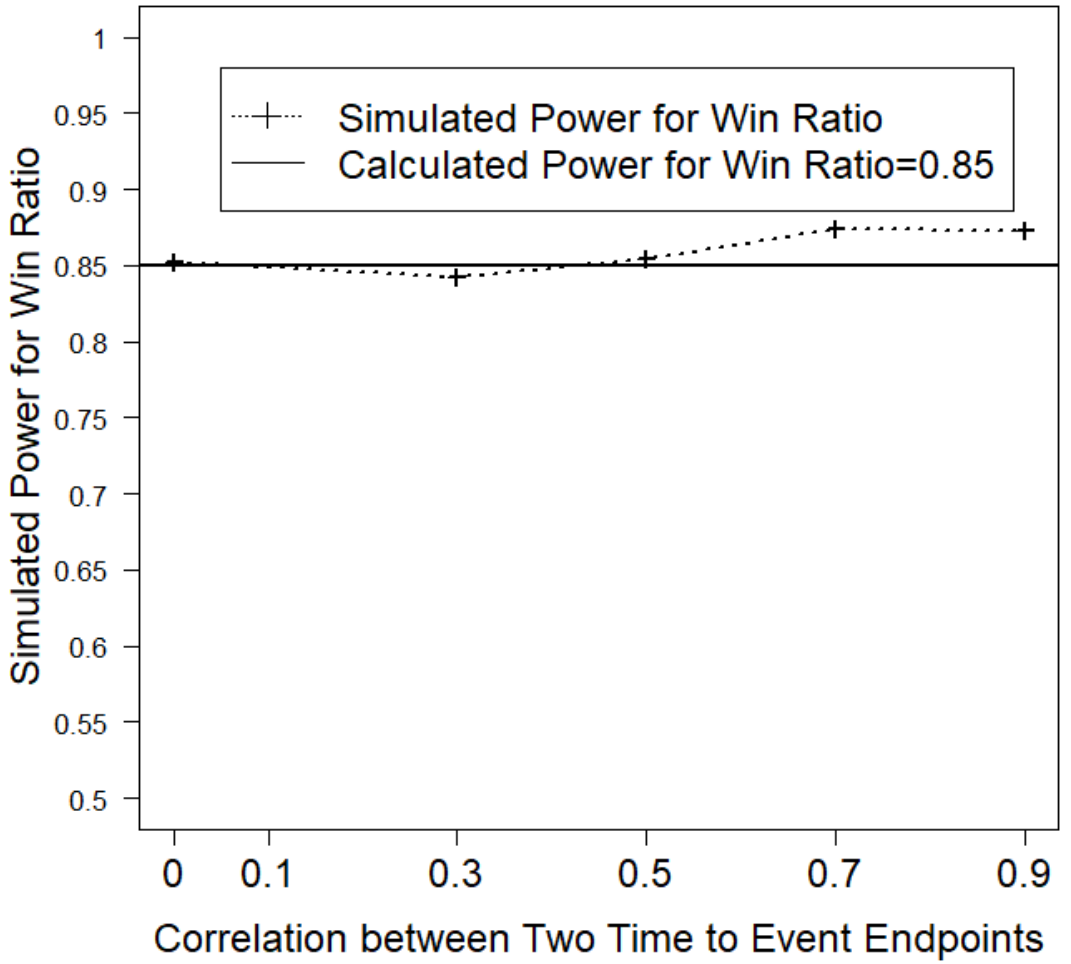
and we may still have

$$\frac{Prob(A\ wins|FU = S)}{Prob\ (B\ wins|FU = S)} \approx \frac{Prob(A\ wins|FU = S, independence)}{Prob\ (B\ wins|FU = S, independence)}.$$

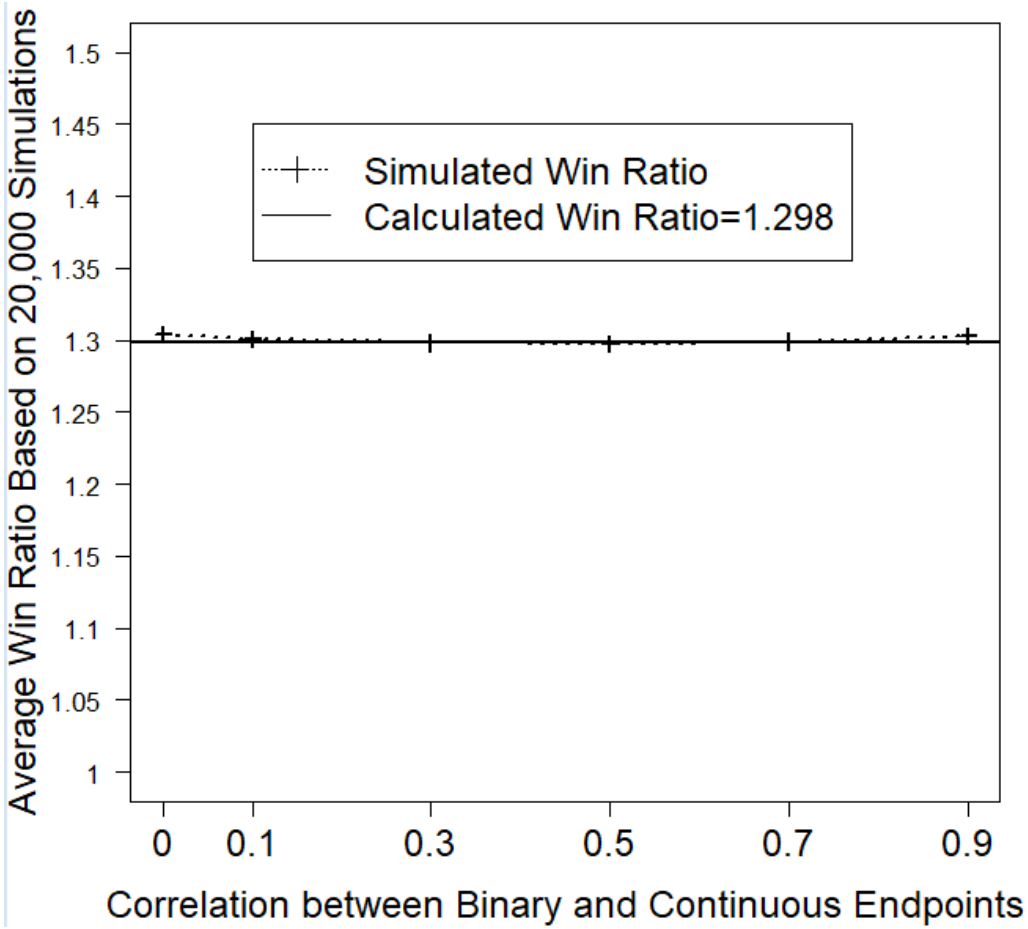
  - The magnitude of win ratio may not change much under various strengths of correlation → may not have much impact on power (checked by 20,000 simulations).
  - The probability of ties may increase slightly as correlation increases → the power may be slightly lower with positive correlation than under independence.

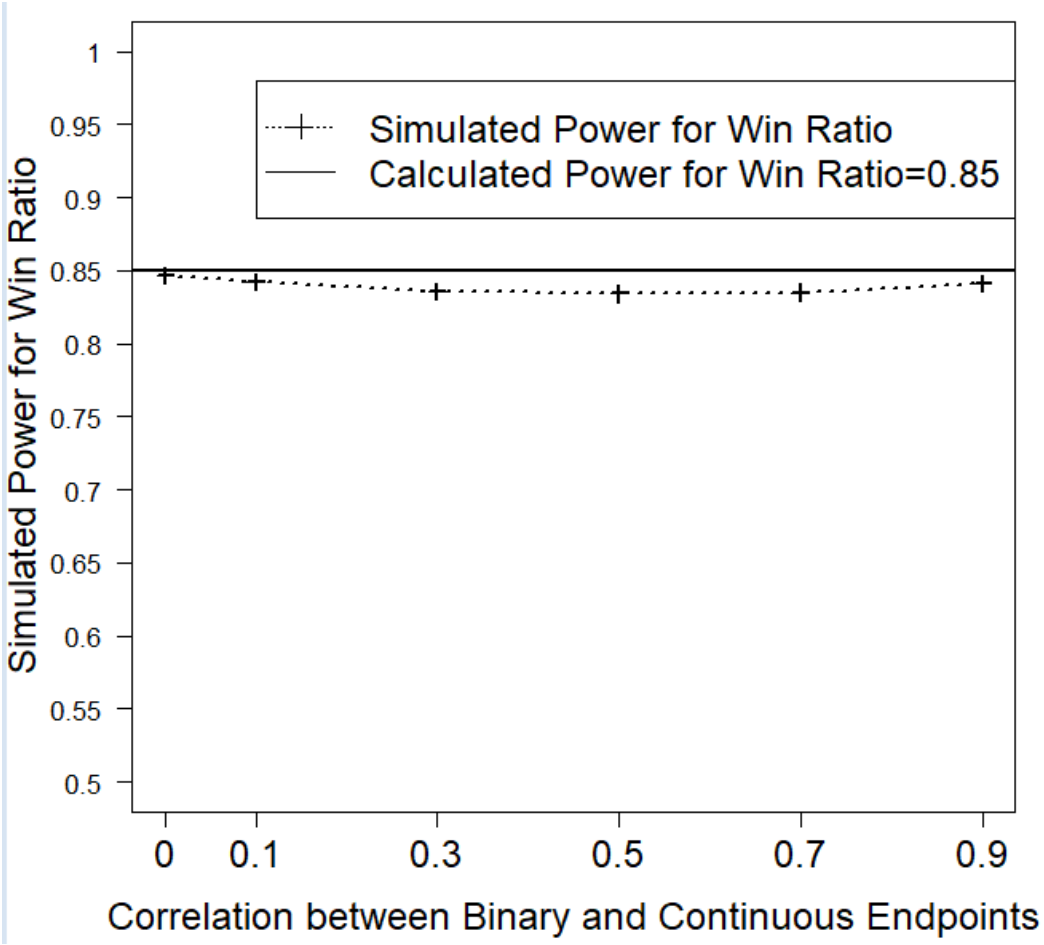
Investigation on Independence Assumption
Two Time to Event Endpoints, 20,000 simulations





Investigation on Independence Assumption
One Binary and One Continuous Endpoints, 20,000 simulations





What if I want to use other win measures in trial design?

- Win Odds
  - Sample size and power formula for WO (Gasparyan et al. 2021) assume $p_{tie} = 0$.
  - It's super conservative if the probability of ties is high

$$\boldsymbol{WO} = \frac{P(\text{A wins on } Y_1) + P(A \text{ wins on } Y_2, \text{ties on } Y_1) + \cdots + P(A \text{ wins on } Y_K, \text{ties on } Y_1 \text{ through } Y_{K-1}) + \frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{P_{tie}}}{P(\text{A loses on } Y_1) + P(A \text{ loses on } Y_2, \text{ties on } Y_1) + \cdots + P(A \text{ loses on } Y_K, \text{ties on } Y_1 \text{ through } Y_{K-1}) + \frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{P_{tie}}}$$

- Net Benefit and DOOR
  - No formula-based sample size and power calculations

$$\boldsymbol{NB} = \boldsymbol{P(A \text{ wins})} - \boldsymbol{P(A \text{ loses})} = NB(S; Y_1) + NB(S; Y_2, \text{ties on } Y_1) + \cdots + NB(S; Y_K, \text{tiese on } Y_1, \ldots, Y_{K-1})$$

$$\boldsymbol{DOOR}(\boldsymbol{S}; \boldsymbol{Y_1}, \ldots, \boldsymbol{Y_K}) = \boldsymbol{P}(\boldsymbol{A \text{ wins on } Y_1}, \ldots, \boldsymbol{Y_K}) + \frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{P}(\boldsymbol{\text{ties}})$$

Our Solution

- We note the relationships between WR, WO, NB and DOOR (in Dong et al., 2023)

$$WO = \frac{WR - \frac{1}{2}p_{tie}(WR-1)}{1 + \frac{1}{2}p_{tie}(WR-1)},$$

$$WO = \frac{1+NB}{1-NB}$$

$$NB = \frac{WR-1}{WR+1}(1-p_{tie}),$$

$$DOOR = \frac{NB+1}{2}$$

Our Solution

- We use the relationships between WR, WO, NE and DOOR to derive new sample size and power formulas
- For Win Odds

$$N_{WO} = \frac{4(1+p_{tie})(1-p_{tie})\left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right)^2}{3r(1-r)\left(\log\left(WO(S;Y_1,\ldots,Y_K)\right)\right)^2}, \quad (3)$$

$$\text{Power}_{\text{WO}} = 1 - \Phi\left(Z_{1-\frac{\alpha}{2}} - abs(\log(WO(S;Y_1,\ldots,Y_K)))\sqrt{\frac{3r(1-r)N}{4(1+p_{tie})(1-p_{tie})}}\right) \quad (4)$$

- Similar formulas for NB and DOOR

Take Home Messages

- Benefits of our new tool:
  - Develop a meaningful and justifiable range of overall win ratio values that are needed for formula-based sample size and power calculations.
  - Quickly evaluate the pros and cons under different set-ups:
    - Tradeoffs on using more or fewer endpoints
    - Tradeoffs on different ranking of endpoints
- Findings with our new tool
  - Correlations between endpoints have minimal impact on the magnitude of win ratio.
- If you're not comfortable with the independence assumption in our new tool, use the tool to produce an initial sample size for simulation studies.

Our New Tool for WO, NB and DOOR

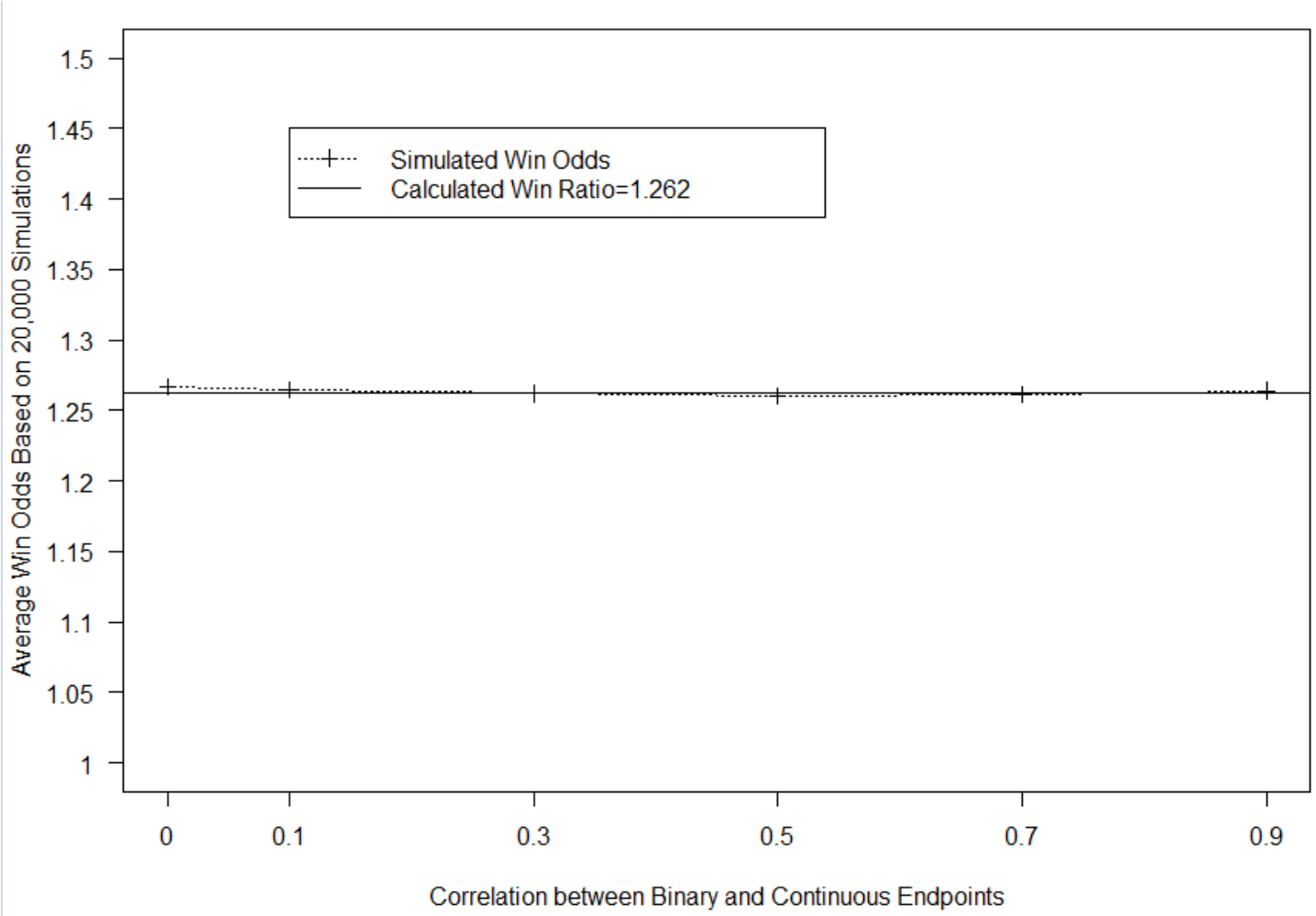
- Assume **INDEPENDENCE** of endpoints
  - WO can be re-written as the weighted sum of marginal WOs plus a term involving probability of ties.

$$\boldsymbol{WO = WO_{Y_1} * \xi_{Y_1} + WO_{Y_2} * \xi_{Y_2} + \cdots + WO_{Y_K} * \xi_{Y_K} + 1 * \in_{(K+1)}}$$

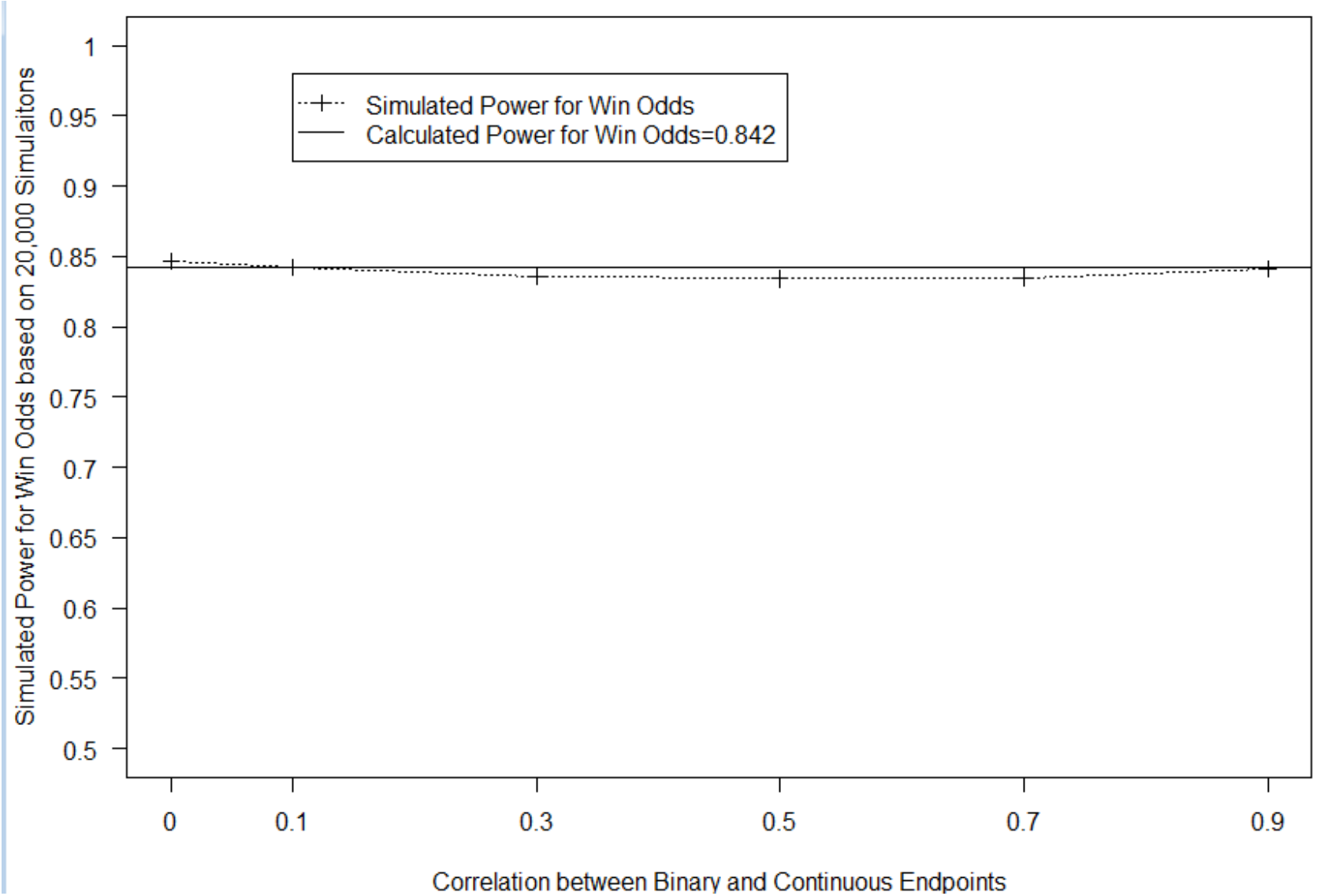
with $\boldsymbol{\xi_{Y_k}} = \frac{WR_{Y_k}\boldsymbol{w_k}}{WO_{Y_k}} * common\ term$

  - Similarly NB and DOOR can be re-written as the weighted sum invovling marginal NBs and DOORs

WO Dependence on Correlation
One Binary and One Continuous Endpoints



Power Dependence on Correlation – WO
One Binary and One Continuous Endpoints



R shiny App and R package

- Accessing R shiny App
  - Through our O2E website: https://dcri.org/o2e
  - Direct link: https://duke-som.shinyapps.io/Hierarchical_Endpoints/
- R Package in CRAN: powerHE

References

1. Barnhart, H. X., Lokhnygina, Y., Matsouaka, R. A., & Rockhold, F. W. (2024). Trial Design with Win Statistics for Multiple Time-to-Event Endpoints with Hierarchy. *Statistics in Biopharmaceutical Research*, 1–24. https://doi.org/10.1080/19466315.2024.2365629
2. Barnhart, H. X., Lokhnygina, Y., Matsouaka, R. A., Halabi, S., Yanez, D., Mentz, R.J., & Rockhold, F. W. (2025) Sample Size and Power Calculations with Win Measures Based on Hierarchical Endpoints. *Statistics in Medicine*, accepted.
3. R Shiny App: https://duke-som.shinyapps.io/Hierarchical_Endpoints/
4. R Package: powerHE

Backup slides for R shiny App Demo

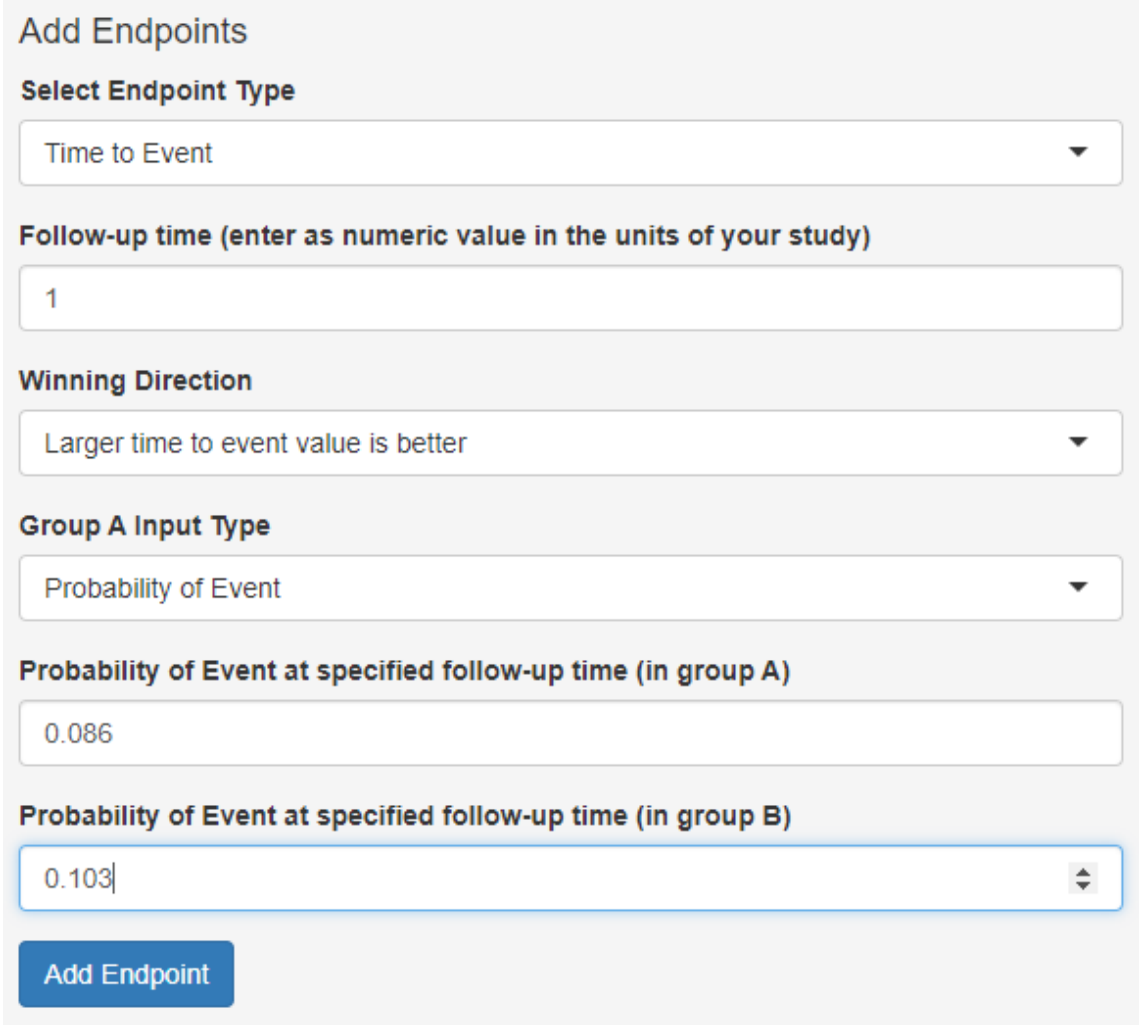
Heart-FID Data for each of the three endpoints

Observed Death Rate at 1 Year		Observed Averaged Number (Adjusted for FU Time) of Hospitalizations for Heart Failures at 1 Year		Mean (SD) of Observed and Imputed Change From Baseline to Month 6 in 6-min Walk Distance	
FCM	Placebo	FCM	Placebo	FCM	Placebo
0.086	0.103	0.257	0.332	-22.22 (106.83)	-24.02 (101.17)

R Shiny App Demo Example: Enter first endpoint

Observed Death Rate at 1 Year

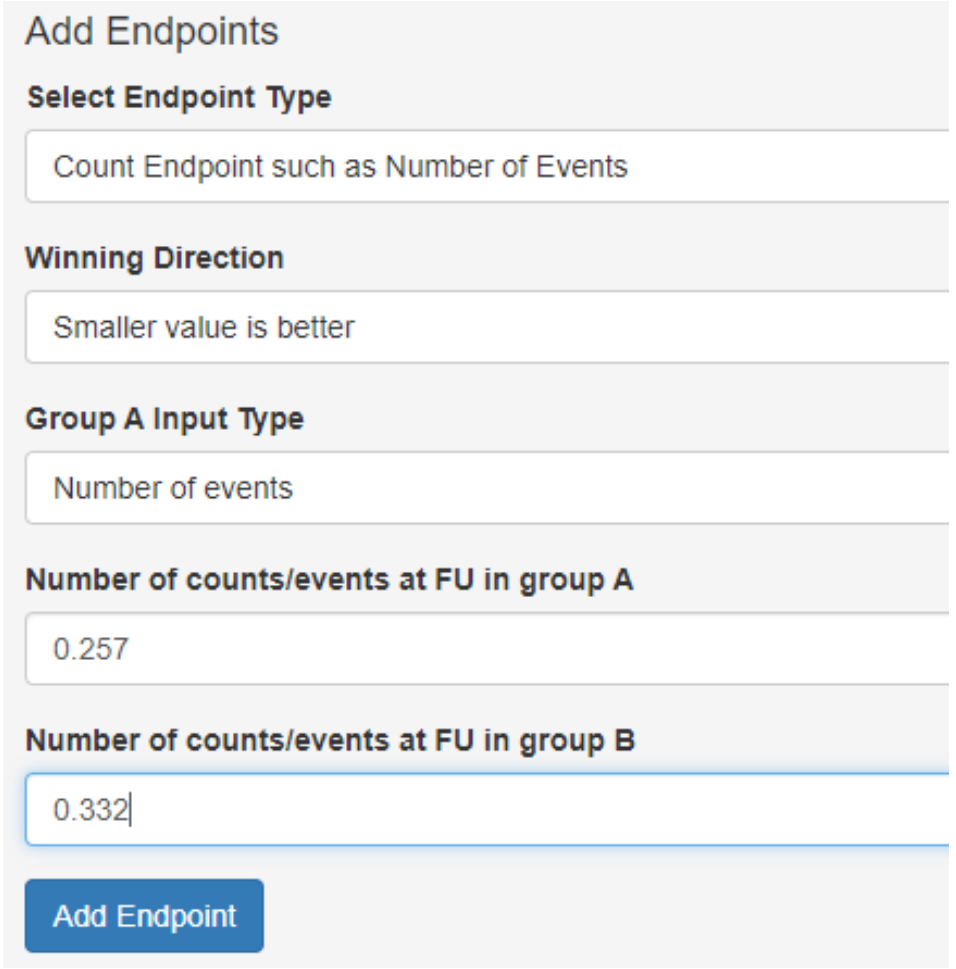
FCM	Placebo
0.086	0.103



R Shiny App Demo Example: Enter the second endpoint

Observed Averaged Number (Adjusted for FU Time) of Hospitalizations for Heart Failures at 1 Year

FCM	Placebo
0.257	0.332



Example 2. Power for sample size=3064 with Two Endpoints

io

Details

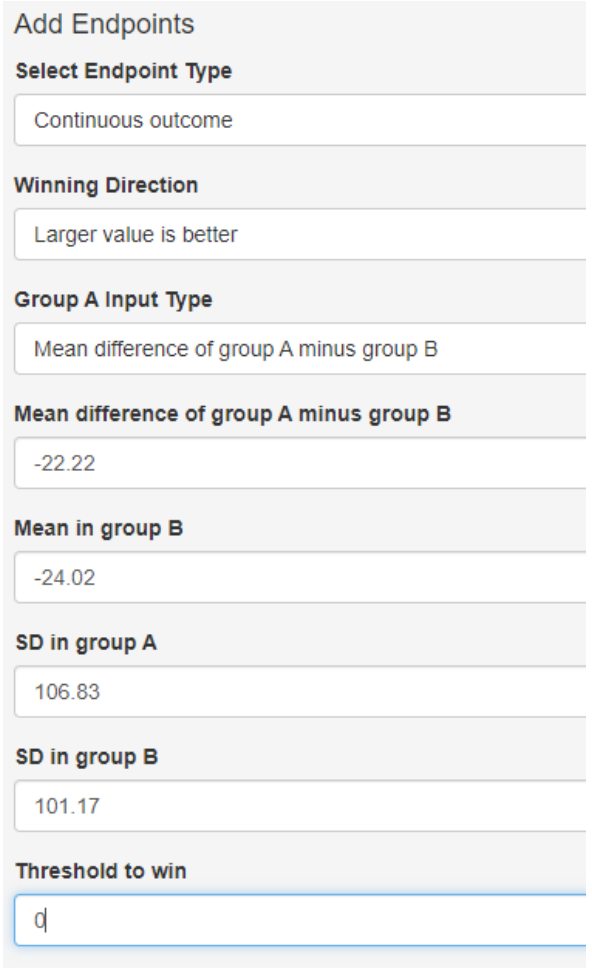
```
 1: TTE
ity of Event at specified follow-up time (in group A) = 0.086
ity of Event at specified follow-up time (in group B) = 0.103
p time = 1
ing Direction = GT
 2: Count
f counts/events at FU in group A = 0.257
f counts/events at FU in group B = 0.332
nning Direction = LT
```

- Power=94.8%

R Shiny App Demo Example: Enter the third endpoint

Mean (SD) of Observed and Imputed Change From Baseline to Month 6 in 6-min Walk Distance

FCM	Placebo
-22.22 (106.83)	-24.02 (101.17)



Example 2. Power for sample size=3064 with Three Endpoints

Power=91.6%

Endpoint Details

```
Endpoint 1: TTE
Probability of Event at specified follow-up time (in group A) = 0.086
Probability of Event at specified follow-up time (in group B) = 0.103
Follow-up time = 1
TTE Winning Direction = GT
Endpoint 2: Count
Number of counts/events at FU in group A = 0.257
Number of counts/events at FU in group B = 0.332
Count Winning Direction = LT
Endpoint 3: Continuous
Mean in group A = -22.22
Mean in group B = -24.02
SD in group A = 106.83
SD in group B = 101.17
Threshold to win = 0
Continuous Winning Direction = GT
```

Example 2. Power for sample size=3064 with Three Endpoints using R package, *powerHE*

```
endpoints_input <- list(
 list(type = "TTE",
    tte.winning.direction = "GT",
    er.a = 0.086,
    er.b = 0.103,
    s=5),
 list(type = "Count",
    count.winning.direction = "LT",
    lam.a = 0.257,
    lam.b = 0.332),
 list(type = "Continuous",
    continuous.winning.direction = "GT",
    mu.a = -22.22,
    mu.b = -24.02,
    sd.a = 106.83,
    sd.b = 101.17,
    delta =0))
```

```
powerHE(endpoints_input,
    sample.size = 3064,
    alpha = 0.05,
    rratio = 0.5,
    output = "WR")
```

```
$p_tie
[1] 0.8198580 0.6032462 0.0000000

$p_tie_overall
[1] 0

$w
[1] 0.1752917 0.2983969 0.5263113

$wr
[1] 1.208783 1.342960 1.019714

$wr_c
[1] 1.149312

$power_c.WR
[1] 0.9155283
```

Take Home Messages

- Benefits of our new tool:
  - Develop a meaningful and justifiable range of win parameter values that are needed for formula-based sample size and power calculations.
  - Quickly evaluate the pros and cons under different set-ups:
    - Tradeoffs on using more or fewer endpoints
    - Tradeoffs on different ranking of endpoints
- Win ratio depends on the follow-up time
- Win ratio does not always win in terms of power
- Adding a less important endpoint for the purpose of breaking ties may not always win in terms of power
- If you're not comfortable with the independence assumption in our new tool, use the tool to come up with an initial sample size for simulation studies.

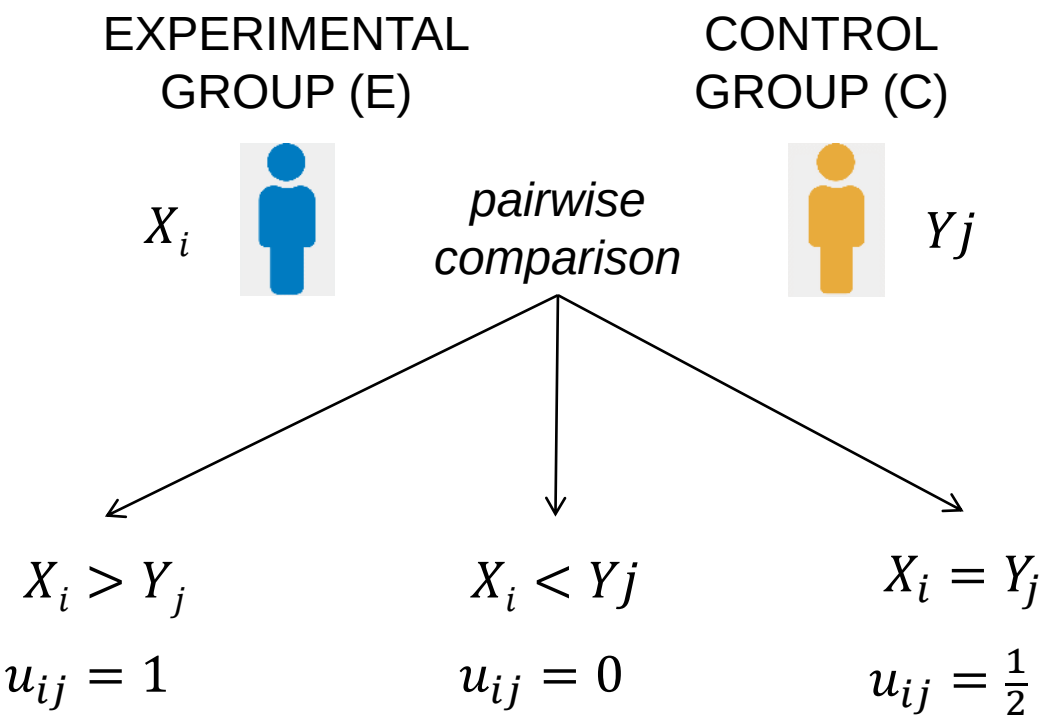
Involving patients in the design of clinical trials using prioritized outcomes

Marc Buyse, ScD
Vancouver, Canada
21 May 2025

Outline

- Wilcoxon-Mann-Whitney test
- Generalized Pairwise Comparisons
- Win Ratio
- Net Treatment Benefit
- Contributions of individual prioritized outcomes
- Challenges in prioritizing outcomes
- A software to elicit patient preferences

Wilcoxon-Mann-Whitney test



Wilcoxon-Mann-Whitney statistic

$$U = \frac{1}{m \cdot n} \sum_{i=1}^{m} \sum_{j=1}^{n} u_{ij}$$

Wilcoxon-Mann-Whitney effect, also called Probabilistic Index (PI)

$$PI := \mathbb{P}(X > Y) + \tfrac{1}{2}\,\mathbb{P}(X = Y)$$

$$\widehat{PI} = U$$

Wilcoxon. Biometrics 1945;1:80. Mann & Whitney. Ann Math Stat 1947;18:50. Hoeffding. Ann Math Stat 1948;19:293

Generalized Pairwise Comparisons

EXPERIMENTAL GROUP (E)

CONTROL GROUP (C)

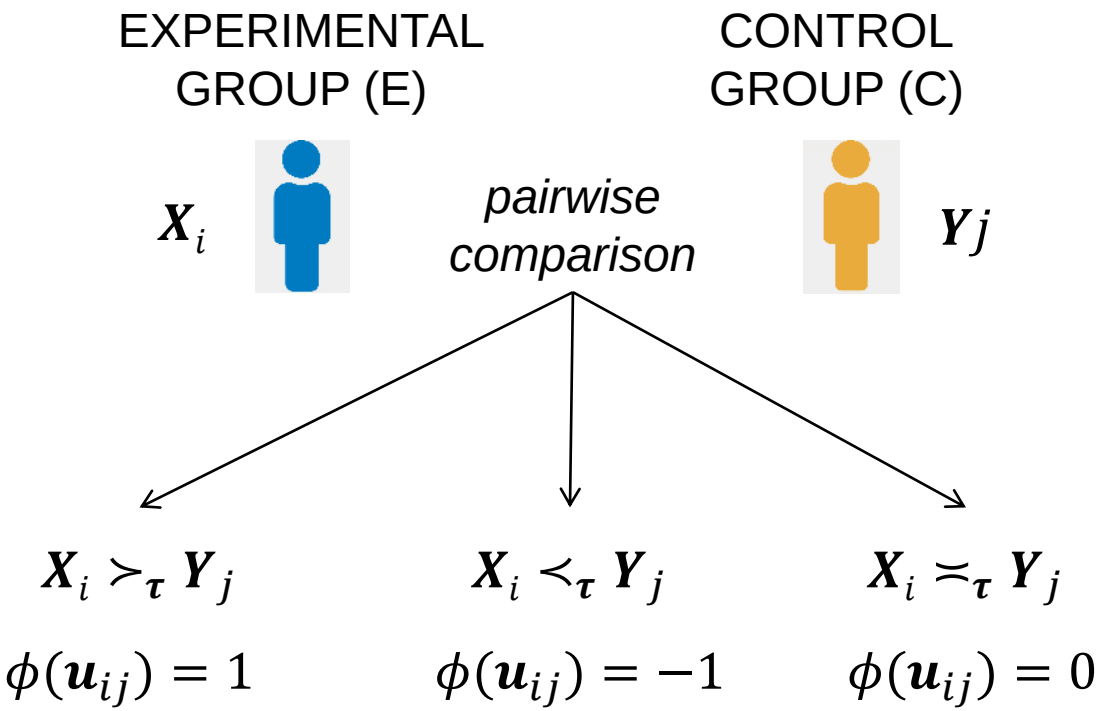
X_i  *pairwise comparison*  Y_j

Generalizations:

- Times to event [1]
- Outcomes of any type [2]
- Multiple outcomes [3]
- Survival + other outcome [4,5]
- Prioritized outcomes [2]
- Thresholds of clinical similarity [2]

[1] Gehan. Biometrika 1965;52:203. [2] Buyse. Stat Med 2010;29:3245. [3] O'Brien. Biometrics 1984;69:1079.
[4] Moyé et al. Stat Med 1992;11:1705. [5] Finkelstein & Schoenfeld. Stat Med 1999;18:1341.

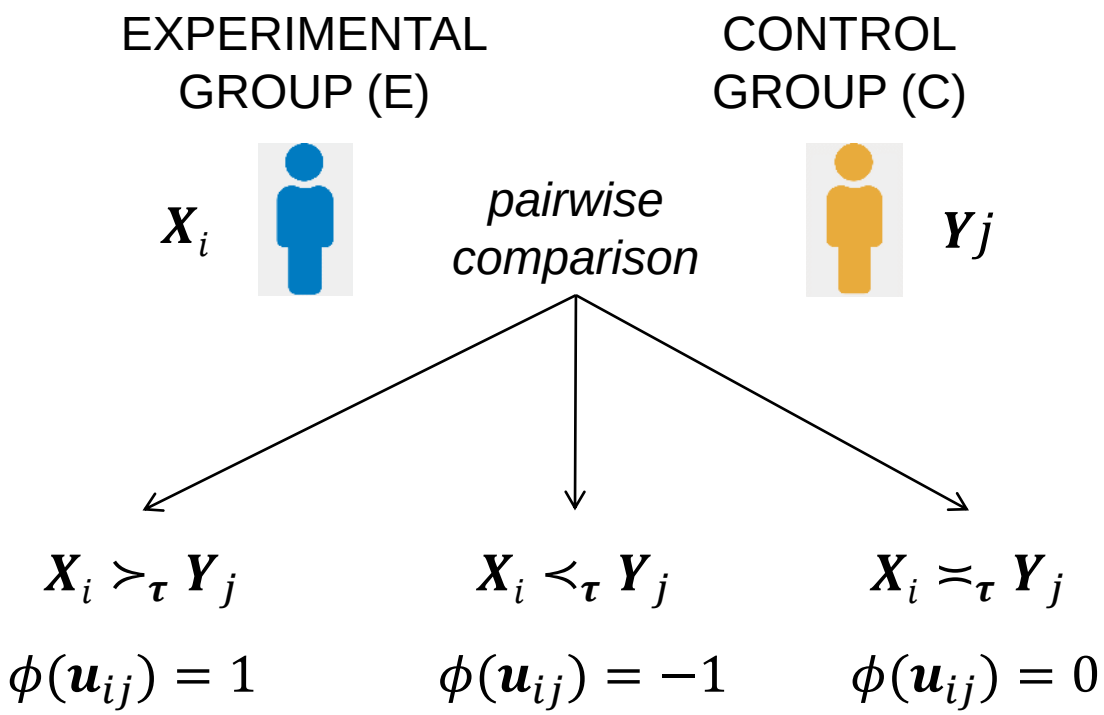
Generalized Pairwise Comparisons



$$U = \frac{1}{m \cdot n} \sum_{i=1}^{m} \sum_{j=1}^{n} \phi(u_{ij})$$

Ramchandani et al. Biometrics 2016;72:926.

Multiple prioritized outcomes



$$U = \frac{1}{m \cdot n} \sum_{i=1}^{m} \sum_{j=1}^{n} \phi(\boldsymbol{u}_{ij})$$

$\phi(\boldsymbol{u}_{ij})$ can be as complex as required

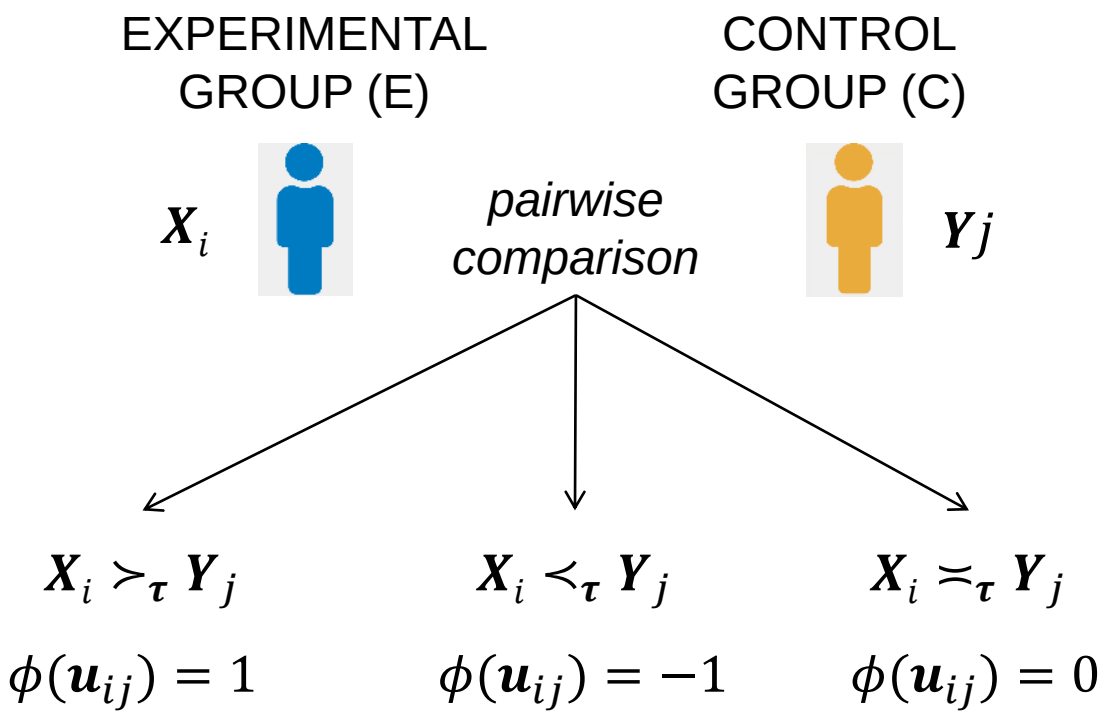
e.g., for d prioritized outcomes

$$\phi(\boldsymbol{u}_{ij}) = u_{ij1} + u_{ij2}\, \mathbb{I}\left[u_{ij1} = 0\right] + \cdots$$

$$+ u_{ijd}\, \mathbb{I}\left[u_{ij1} = 0 \wedge \ \ldots \wedge u_{ij(d-1)} = 0\right]$$

Ramchandani et al. Biometrics 2016;72:926. Buyse. Stat Med 2010;29:3245.

Net Treatment Benefit (NTB)



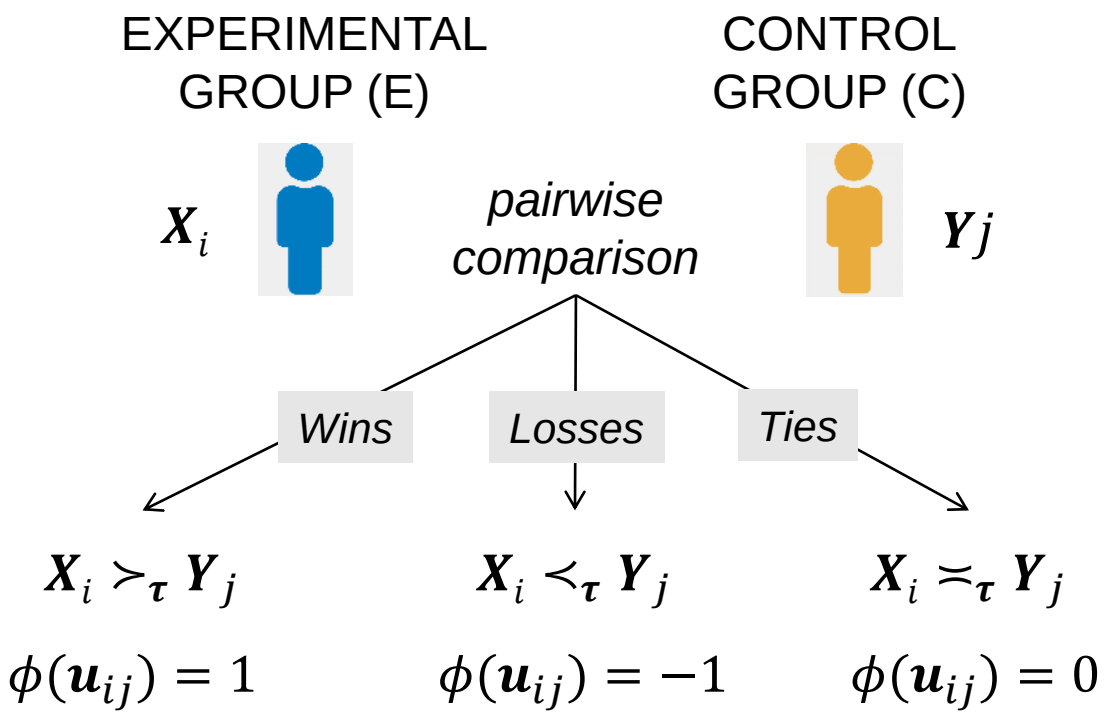
$$U = \frac{1}{m \cdot n} \sum_{i=1}^{m} \sum_{j=1}^{n} \phi(u_{ij})$$

Net Treatment Benefit (NTB)

$$NTB := \mathbb{P}(X_i \succ_\tau Y_j) - \mathbb{P}(X_i \prec_\tau Y_j)$$

Buyse. Stat Med 2010;29:3245.

Win Ratio (WR)



$$U = \frac{1}{m \cdot n} \sum_{i=1}^{m} \sum_{j=1}^{n} \phi(u_{ij})$$

Net Treatment Benefit (NTB)

$$NTB := \mathbb{P}(X_i \succ_\tau Y_j) - \mathbb{P}(X_i \prec_\tau Y_j)$$

Win Ratio (WR)

$$WR := \frac{\mathbb{P}(X_i \succ_\tau Y_j)}{\mathbb{P}(X_i \prec_\tau Y_j)}$$

Pocock et al. Eur Heart J 2012;33:176.

Interpretation of Win Ratio

WR is the odds ratio for a dichotomous outcome.

For a time to event under proportional hazards, WR is the reciprocal of the hazard ratio (HR)

$$WR = \frac{1}{HR}$$

But how about multiple outcomes?

Oakes. Biometrika 2016;103:742.

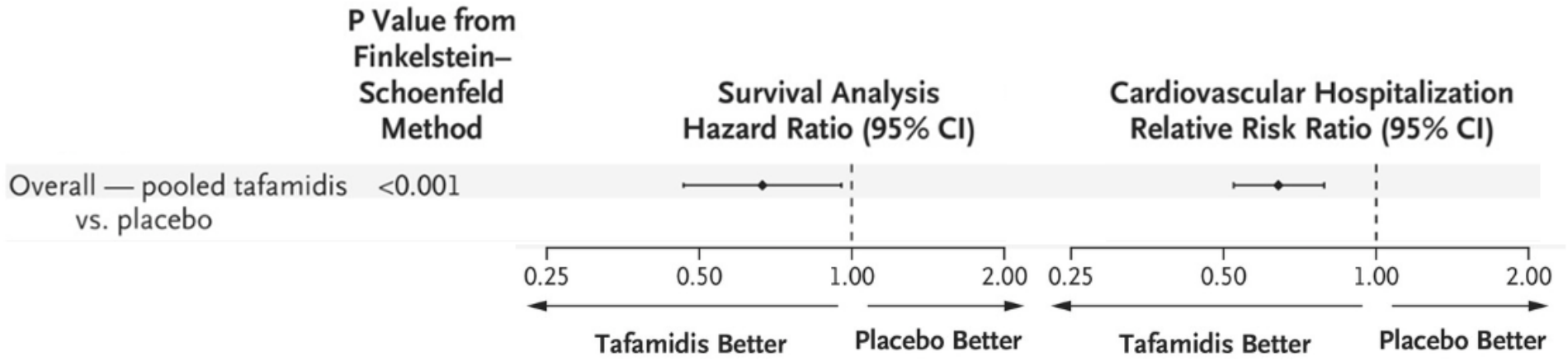
The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 SEPTEMBER 13, 2018 VOL. 379 NO. 11

Tafamidis Treatment for Patients with Transthyretin Amyloid Cardiomyopathy

GPC with two prioritized outcomes :

1. survival
2. frequency of cardiovascular-related hospitalizations



The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 SEPTEMBER 13, 2018 VOL. 379 NO. 11

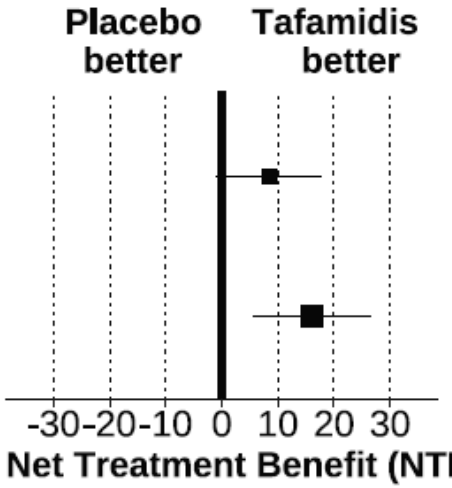
Tafamidis Treatment for Patients with Transthyretin Amyloid Cardiomyopathy

The win ratio[24] (number of pairs of treated-patient "wins" divided by number of pairs of placebo-patient "wins") may be helpful in interpreting the Finkelstein–Schoenfeld result. The win ratio is 1.695 (95% confidence interval [CI], 1.255 to 2.289).

Interpretation ?

Net Treatment Benefit

Prioritized Outcomes	Total Pairs	Neutral Pairs	Favor. Pairs	Unfavor. Pairs	IP [1] (cumul)	IC [2] (cumul)	NTB (95% CI)	WR [3] (95% CI)	*p* – value
All-Cause Mortality	100.0	43.1	32.7	24.2	56.9 (56.9)	8.5 (8.5)	8.5 (-0.9,17.6)	1.35 (0.97,1.87)	0.075
CV-Related Hospitalization	43.1	16.8	17.0	9.2	26.2 (83.1)	7.8 (16.2)	16.2 (5.6,26.5)	1.49 (1.15,1.92)	0.0029



All numbers are % except WR and *p*-values

[1] Information Proportion (IP) = Wins + Losses = Favorable Pairs + Unfavorable Pairs

[2] Individual Contribution (IC) = Wins – Losses = Favorable Pairs – Unfavorable Pairs

[3] Win Ratio : Wins / Losses = Favorable Pairs / Unfavorable Pairs

Buyse, Verbeeck, De Backer, Deltuvaite-Thomas, Saad, Molenberghs (Editors). Handbook of Generalized Pairwise Comparisons. Taylor & Francis, 2025.

Interpretation of Net Treatment Benefit

NTB generalizes the difference between the probabilities of success $P^E - P^C$:

$$NTB := \mathbb{P}\left(\boldsymbol{X}_i \succ_{\boldsymbol{\tau}} \boldsymbol{Y}_j\right) - \mathbb{P}\left(\boldsymbol{X}_i \prec_{\boldsymbol{\tau}} \boldsymbol{Y}_j\right)$$

NTB is the difference between the probability that a random patient in the Experimental group has a better outcome than a random patient in the Control group, minus the probability of the opposite situation.

The definition of « a better outcome » can be made as complex as needed clinically.

NTB ranges from -1 to +1, with 0 indicating no treatment effect.

Buyse. Stat Med 2010;29:3245.

Challenges in Prioritizing Outcomes

Software to

- select an appropriate hierarchy of outcomes
- ensure alignment with patient (or clinical expert) priorities
- determine thresholds of clinical similarity

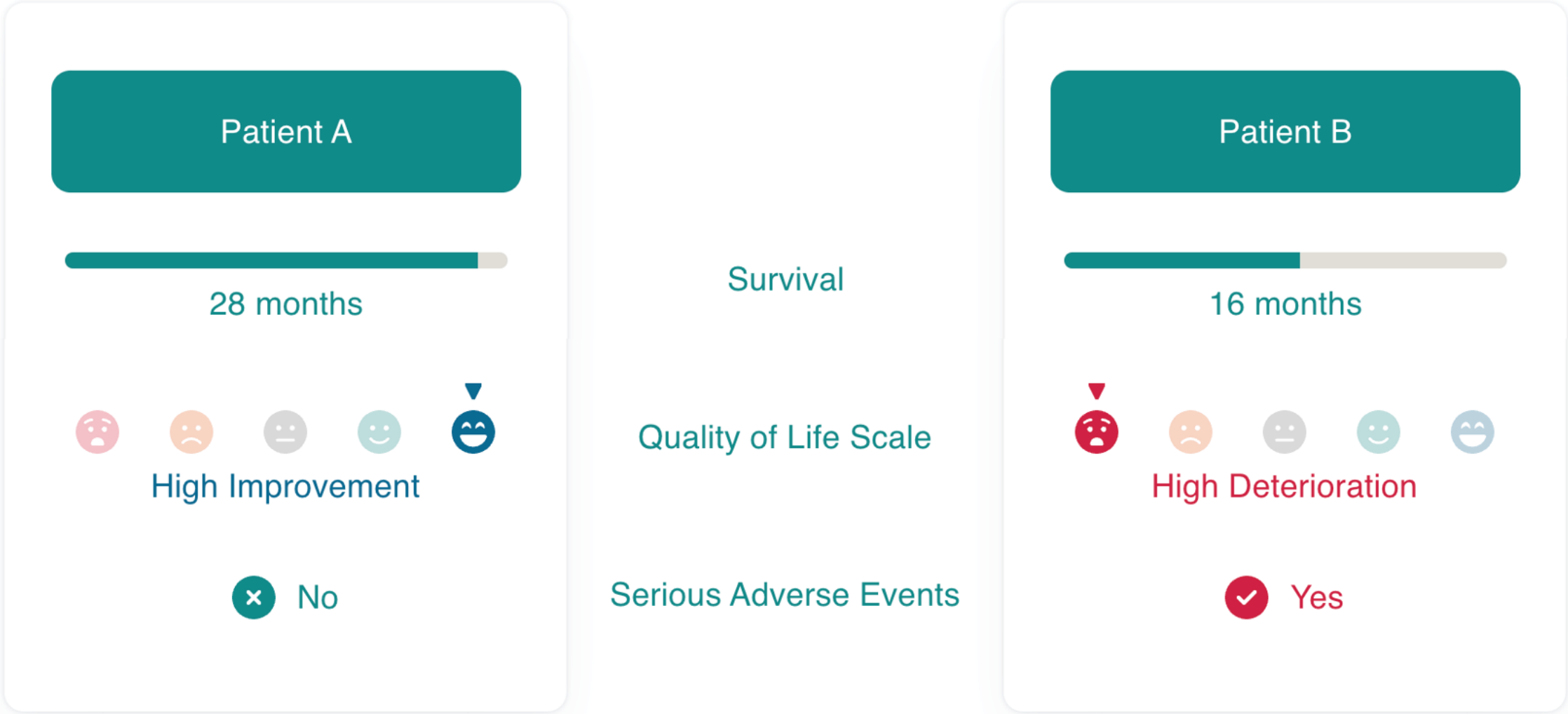
Early engagement of patients / clinicians in the trial design to :

1. identify outcome priorities
2. define clinical thresholds for outcomes where relevant

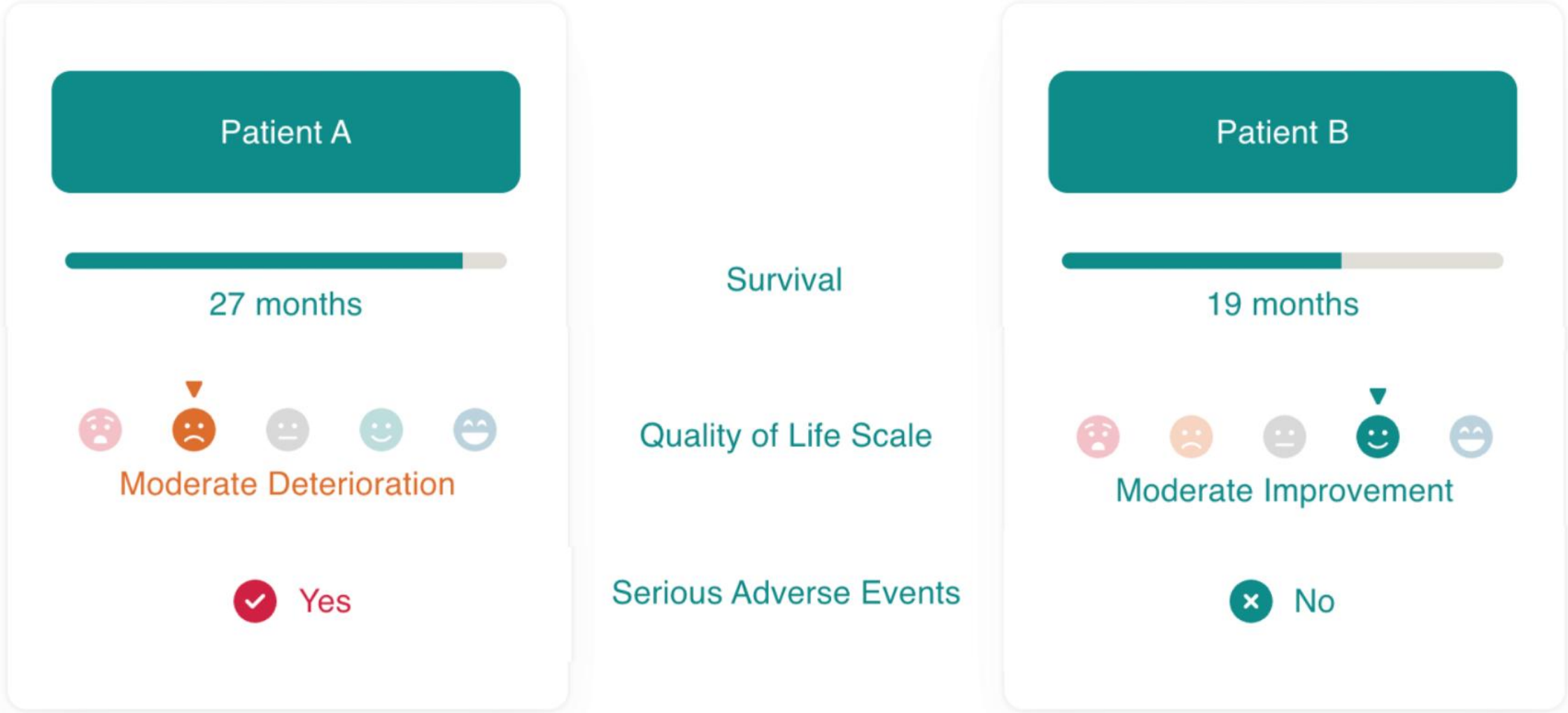
Pairwise comparisons as an intuitive way to elicit choices

"Which of these two scenarios do you prefer (or find more favorable)?"

Pairwise comparisons as an intuitive way to elicit choices



Pairwise comparisons as an intuitive way to elicit choices



Eliciting choices for each respondent

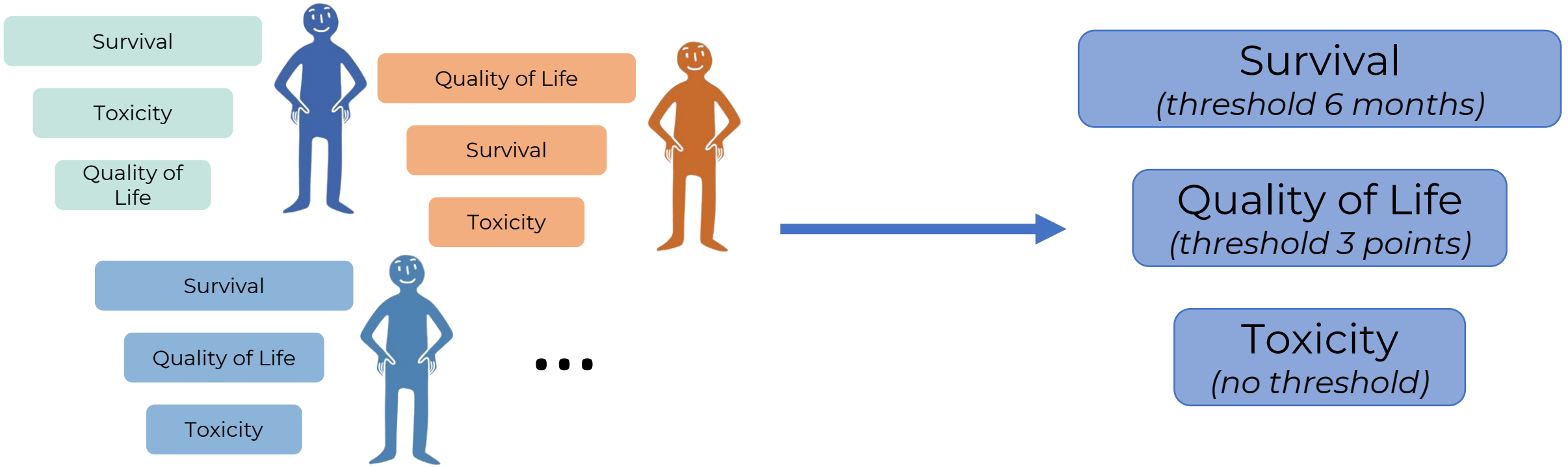
Learn and adapt
- Learn priority
- Adapt trade-off

Select list of outcomes → Generate pairs with outcome values → Pairwise comparisons → Validation → Results

List of prioritized outcomes and thresholds

Aggregating choices of multiple respondents

Aim: to produce a list of prioritized outcomes and thresholds that maximizes agreement across all respondents



Chapman & Hall/CRC

Handbooks of Modern Statistical Methods

"This book stands as a guiding beacon for developers, researchers, and regulators, sparking the evolution of fit-for-dossier trials into agile studies tailored for informed decisions."
Francesco Pignatti, *Head of the Office of Oncology and Haematology, European Medicines Agency, Amsterdam, the Netherlands*

"The editors of this book and the chapter authors are to be commended for consolidating the considerable advances in GPC statistical methods into a single comprehensive resource that should serve as a standard for many years to come."
Gene Pennello, *Mathematical Statistician, US Food and Drug Administration, Bethesda, MD.*

In today's healthcare landscape, there is a pressing need for quantitative methodologies that include the patients' perspective in any treatment decision.

Handbook of Generalized Pairwise Comparisons: Methods for Patient-Centric Analysis provides a comprehensive overview of an innovative and powerful statistical methodology that generalizes the traditional Wilcoxon-Mann-Whitney test by extending it to any number of outcomes of any type and including thresholds of clinical relevance into a single, multidimensional evaluation.

The book covers the statistical foundations of generalized pairwise comparisons (GPC), applications in various disease areas, implications for regulatory approvals and benefit-risk analyses, and considerations for patient-centricity in clinical research. With contributions from leading experts in the field, this book stands as an essential resource for a more holistic and patient-centric assessment of treatment effects.

CRC Press
Taylor & Francis Group
www.routledge.com
CRC Press titles are available as eBook editions in a range of digital formats

an informa business
978-1-032-48805-9

Marc Buyse
Johan Verbeeck
Mickaël De Backer
Vaiva Deltuvaite-Thomas
Everardo D. Saad
Geert Molenberghs

Handbook of Generalized Pairwise Comparisons

Chapman & Hall/CRC
Handbooks of Modern Statistical Methods

Handbook of Generalized Pairwise Comparisons

Methods for Patient-Centric Analysis

Edited by
Marc Buyse
Johan Verbeeck
Mickaël De Backer
Vaiva Deltuvaite-Thomas
Everardo D. Saad
Geert Molenberghs

A Chapman & Hall Book

CRC Press
Taylor & Francis Group

Acknowledgments

JC Chiem, PhD

Sarah Kosta, PhD

Samuel Salvaggio, PhD

Questions

marc.buyse@one2treat.com

Milken Institute School of Public Health
THE GEORGE WASHINGTON UNIVERSITY
THE BIOSTATISTICS CENTER

SCT
SOCIETY FOR CLINICAL TRIALS

The SCT 46th Annual Meeting
Vancouver, Canada,
May 18-21, 2025

The Desirability Of Outcome Ranking

The DOOR to Patient-Centric Benefit-Risk Evaluation

Toshimitsu Hamasaki, PhD, MS, Pstat®
Scott R. Evans, PhD, MS

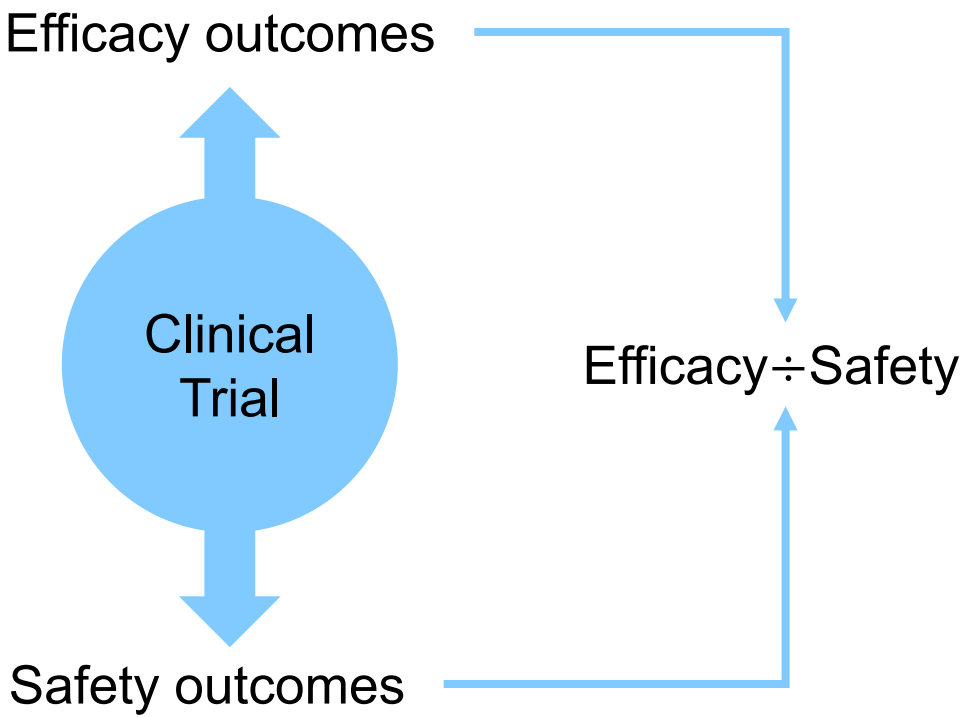
The Biostatistics Center | Department of Biostatistics and Bioinformatics
Milken Institute School of Public Heath | The George Washington University

Research reported in this presentation was supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health(NIH) under Award Number UM1AI104681. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH

Challenges in Benefit:risk Evaluation

Motivations

"**Typical**" benefit:risk analyses



- Compare interventions for each efficacy and safety outcome marginally
- Combine effects formally or informally

- Do tell us patients' journey- what is the most useful information when treating patients?
  - Tradeoff relationship between outcomes
  - Patient experience- the cumulative nature of outcomes on individuals
- Suffer from competing risk complexities during interpretation of individual outcomes
- generalizability? Ex. Efficacy analysis on Intention-to-treat population; safety analysis on safety population

The Desirability Of Outcome Ranking (DOOR) Methodology

Patient-centric evaluation

- A paradigm for the design, monitoring, analysis, interpretation and reporting of clinical trials and other research studies based on patient-centric benefit:risk evaluation (Evans et al. 2015; Evans and Follmann 2016)- **Using Outcomes to Analyze Patients** rather than Patients to Analyze Outcomes
- Implementation of the DOOR methodology in clinical trials
  - ❑ DOOR outcome
  - ❑ DOOR Analyses
    - ➢ Rank-based analysis approach
    - ➢ Grade-based analysis approach
  - ❑ Sample size calculation

Evans SR et al. Clin Infect Dis 2015; 61:800-806. Evans SR, Follmann D. Stat Biopharm Res 2016; 8:386-393

DOOR outcomes

Installing a DOOR outcome

- Requires evaluating the tradeoffs among outcomes, and the cumulative nature of benefits and harms on patients
- Define gradations of patient response to enable recognition of important differences in ultimate responses resulting from therapeutic intervention.
- Determine combination of clinical outcomes that is most influential in selecting treatment/treatment strategy for patients and guide the construction of DOOR outcomes
- How?
  - ❑ *Staphylococcus aureus* bloodstream infection: Survey expert clinicians and patients, and conduct conjoint analysis (Doernberg SB et al. Clin Infect Dis 2019; 68:1691-1698) https://arlg.org/desirability-of-outcome-ranking-door/
  - ❑ Periprosthetic joint infection: Delphi method (Johns BP et al. J Bone Joint Infect 2022; 7:221–229)

DOOR outcomes

ARLG proposed DOOR outcomes

ARLG
Antibacterial Resistance Leadership Group

The Antibacterial Resistance Leadership Group (ARLG) (https://arlg.org/)

- hosts an Innovation Working Group along with colleagues associated with FDA DOOR Fellowships
- developed the DOOR outcomes
  - Acute bacterial skin and skin structure infections (ABSSSI)
  - Bacteremia
  - Complicated intra-abdominal infections (cIAI)
  - Complicated urinary tract infection (cUTI)
  - Hospital-acquired bacterial pneumonia /ventilator-associated bacterial pneumonia (HABP/VABP)

DOOR Outcomes Tailored for Diseases

ARLG proposed DOOR outcomes

- Absence of Clinical Response
- Non- Fatal Serious Adverse Events
- **Infections complications**
- Death

Disease	Infectious Complications
ABSSSI	Unplanned surgical for progression/ complication of original infection; Bacteremia; Septic shock; Osteomyelitis; *c.diff*
Bacteremia	Septic shock; Prolonged bacteremia on Day 5; Supportive complications or monastic site(s) of infection; *c.diff*
cIAI	Bacteremia; Septic shock; Peritonitis; Unplanned surgical for progression/ complication of original infection; *c.diff*
cUTI	Renal or intra-abdominal abscess; Septic shock; Bacteremia; Unplanned surgical for progression/ complication of original infection; *c.diff*
HABP/VABP	Complicated pleural effusion; Lung abscess/necrotizing pneumonia; ARDS; Meningitis; Bacteremia; Septic shock; Need for intubation; *c.diff*

Howard-Anderson J et al. Clin Infect Dis 2024;78:259-268. Kinamon T, et al. Clin Infect Dis 2024; 79:60-69. Kinamon T et al. Clin Infect Dis 2023;77:649-656. Howard-Anderson J et al. Clin Infect Dis 2023;76:e1157-e1165

Applying the DOOR in a Clinical Trial

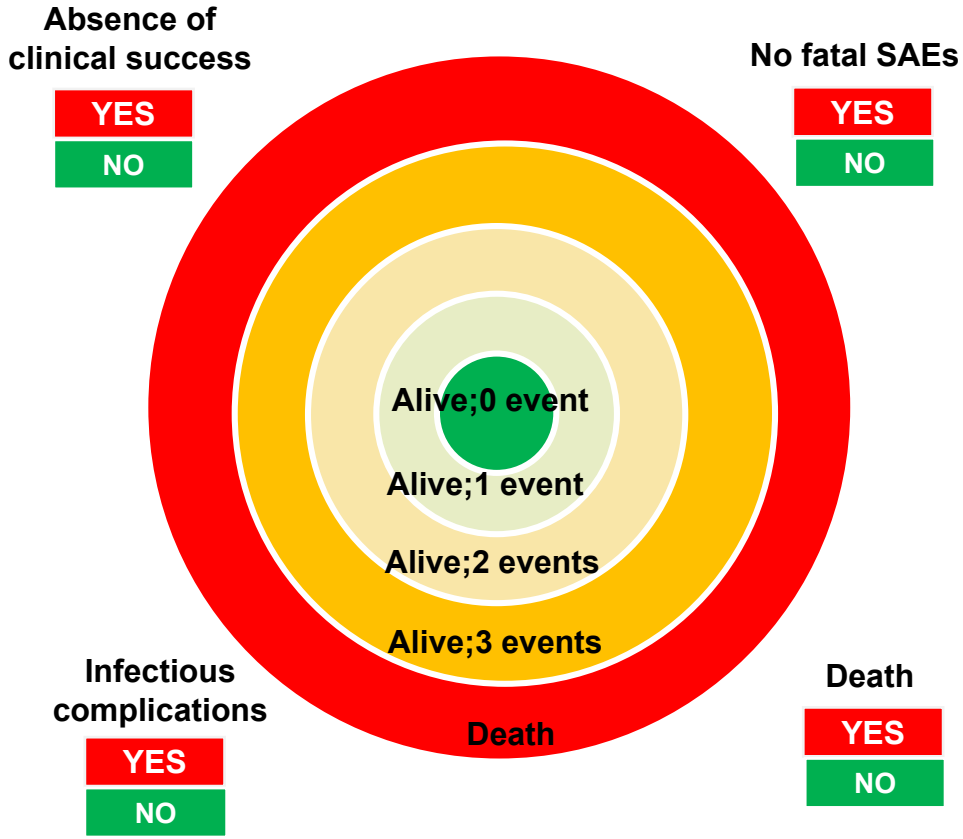
DORI-05: doripenem vs. levofloxacin

DOOR outcome category	Doripenem		Levofloxacin	
	Freq	Prop (%)	Freq	Prop (%)
Alive with no events	263	70.3	253	67.6
Alive with 1 event	93	24.9	111	29.7
Alive with 2 events	16	4.3	9	2.4
Alive with 3 events	1	0.3	1	0.3
Death	1	0.3	0	0
Total	374	100	374	100

From a randomized double-blind clinical trial that evaluated whether intravenous (IV) administration of doripenem (DORI) was inferior to IV administration of levofloxacin in patients with cUTI (complicated urinary tract infection) (Naber KG et al. Antimicrob Agents Chemother 2009; 53:3782-3792) (Howard-Anderson J at. Clin Infect Dis. 2023; 76:e1157-e1165)

ARLG
Antibacterial Resistance Leadership Group

DOOR for cUTI



Renal or intraabdominal abscess; Septic shock; Bacteremia; Recurrent UTI or pyelonephritis; *C. difficile*

Statistical Concerns

Ordinal outcomes analysis

Analysis		Concern
Responder analysis	● Dichotomize to a binary outcome i.e., responder vs non-responder. ● Estimate odds ratio of responder vs non-responder between groups; conduct associated hypothesis testing for the odds ratio	● Loss of information and power by ignoring finer but important gradations of patient status ● Robustness: assumption-reliant (e.g., the model linearity) ● Interpretations are not intuitive
Proportional odds regression	● Estimate common odds ratio by proportional odds regression; Conduct associated methods for hypothesis testing and interval estimation	● Robustness: assumption-reliant (e.g., the model linearity, proportional odds) ● Interpretations are not intuitive

Necessity to Use Absolute Risk Scale

Challenges in use of relative risk

- So how is RR=2 interpreted?
  - Risk of death increases from 1 in 10 to 2 in 10→ RR=2. Very important!
  - Risk of death increases from 1 in 1,000 to 2 in 1,000 → RR=2. Nearly irrelevant.
- Additional challenges arise when interpreting multiple relative risks simultaneously
  - Suppose that with use of an intervention, efficacy doubles: RR=2
  - Further suppose that an equally important safety events double: RR=2
  - This may lead to the belief that there is an equal tradeoff
  - But if the doubling of efficacy is 1 in 10 to 2 in 10 and the doubling of the safety event is 1 in 1,000 to 2 in 1,000, then the tradeoff is unequal
- Interpreting relative risk / ratio measures from multiple outcomes is misleading due to different baseline risks
- Absolute risks summaries are necessary when synthesizing results of composite since component outcomes must be independently evaluated but interpreted simultaneously

The DOOR Methodology

Key principles

Sensitivity

- Simple tools for assessing the robustness of results

Robustness

- Avoid reliance on strong and unconfirmable assumptions

Unbiased estimators

- Avoidance of over or under-estimation of treatment effects

Simple implementation

- No advanced mathematical skills or programming knowledge required

KEY PRINCIPLES

Error controls

- Avoidance of false positive or false negative results

Generalizability

- Clearer estimands and population

Intuitive reporting & presentation

- Enhanced understanding, informed decision-making, better communication with patients

Intuitive measures

- A clear and easily interpretable summary measure

Analytical Methods

Two analyses of DOOR outcomes

Rank-based analysis	Grade-based analysis
❑ The DOOR probability ➢ A patient randomly selected from one group has a more desirable outcome than a patient randomly selected from the other group- 50% if two DOOR outcomes are identical between groups ➢ Population causal effect, not individual causal effect – Not depend on the specific potential outcome pairings (Fay et al 2018) ➢ Pairwise comparisons at individual patient level- Estimated by Wilcoxon-Mann-Whitney (WMW) statistic	❑ Partial credit keys ➢ Evaluation of the relative importance of DOOR outcome categories: robustness analyses for the DOOR probability-based analysis ➢ Methods for continuous outcomes as if they were continuous after assigning grading keys- Welch's t-statistic based method

Fay M et al. Stat Med. 2018;37:2923-2937

Analytical Methods

Recommended statistical analysis plan

Analysis	Outcome	Statistical method
Descriptive analysis	• DOOR • Components	• Summary distribution table by intervention group • Bar-chart by intervention group
	• DOOR and Components	• Anthology of Patient Stories (APS) plot
Rank-based analysis: DOOR probability	• DOOR • Components	• Forest Plot of estimates of the DOOR probability for the DOOR outcome and respective components
	• DOOR	• Forest plot of the estimates for the cumulative DOOR probability based on sequential dichotomization of the DOOR outcome
Grade-based Analysis: Partial Credit	• DOOR	• Welch's t-statistic based analysis • Scatter plot of the differences in mean partial credit between interventions vs the corresponding DOOR probabilities

Implementation of the DOOR analyses

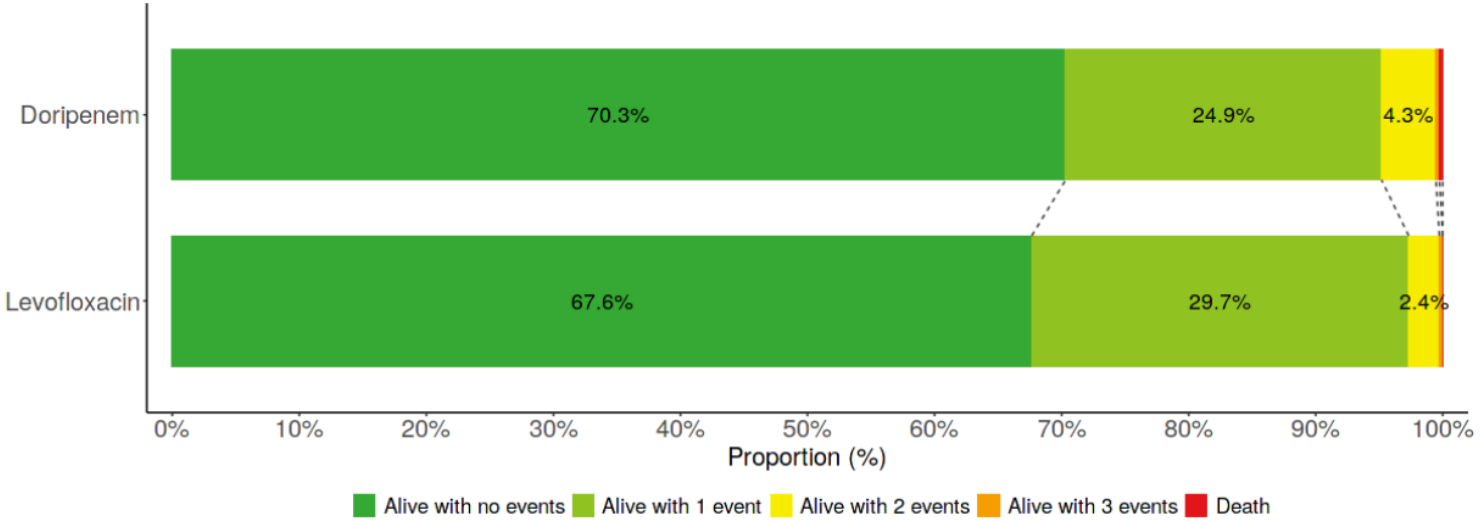
DOOR analysis online tools

	Standard Edition	Professional Edition
Data Input	Summary table by group	Individual patient-level data
Analysis		
1. Descriptive analysis Summary table Bar-chart Anthology of patient stories plot	✓ ✓	✓ ✓ ✓
2. Rank-based analysis DOOR pro forest plot Dichotomized DOOR prob forest plot	✓ ✓	✓ ✓
3. Grade-based analysis Partial credit analysis summary Partial credit vs DOOR prob plot Partial credit forest plot	✓ ✓ ✓	✓ ✓ ✓
4. Tie-breaker analysis		✓
5. Inverse probability weighting		✓
Labels customization, Data save	✓	✓

Applying the DOOR in a Clinical Trial

DORI-05: Descriptive statistics

	Doripenem				Levofloxacin					
			Cumulative				Cumulative		Expected Gained (+) or Loss (-) (per1000)	
DOOR	n	(%)	n	(%)	n	(%)	n	(%)	Per Category	Cumulative
Alive with no events	263	70.3	263	70.3	253	67.6	253	67.6	27	27
Alive with 1 event	93	24.9	356	95.2	111	29.7	364	97.3	-48	-21
Alive with 2 events	16	4.3	372	99.5	9	2.4	373	99.7	19	-3
Alive with 3 events	1	0.3	373	99.7	1	0.3	374	100.0	0	-3
Death	1	0.3	374	100.0	0	0.0	374	100.0	3	0
Total (N)	**374**				**374**					

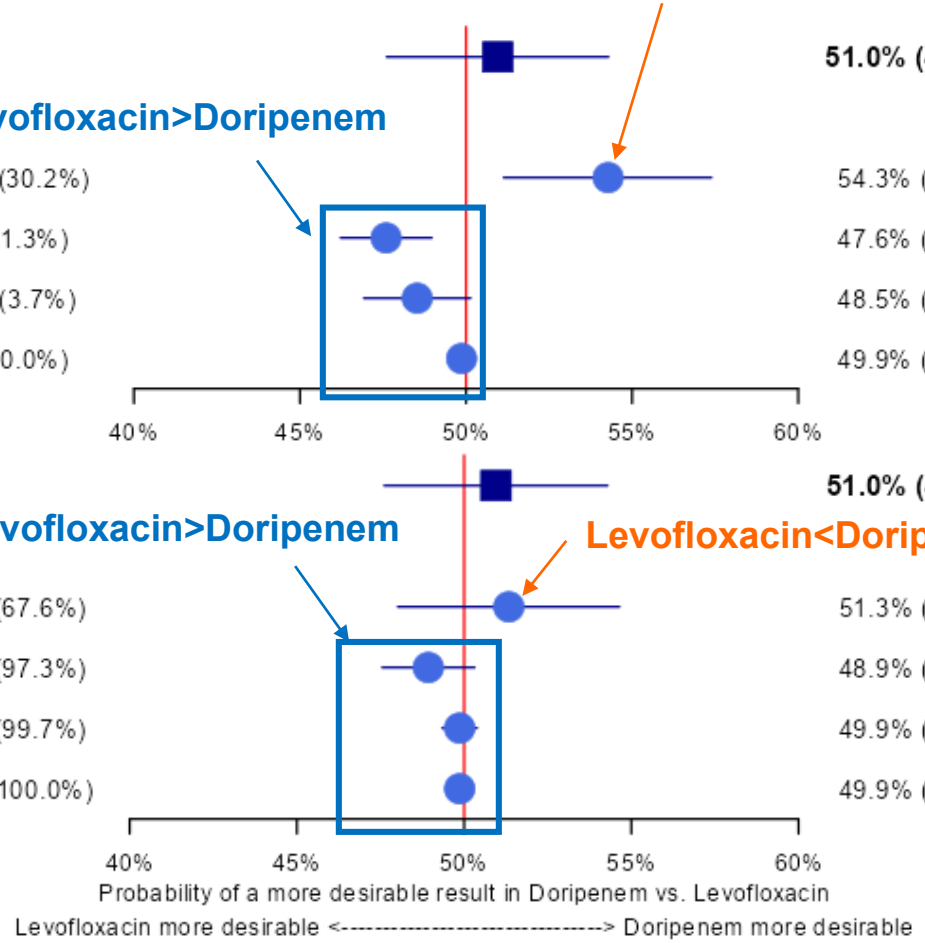


Howard-Anderson J et al. Clin Infect Dis 2023 76:1157-1165.

Applying the DOOR in a Clinical Trial

DORI-05: Rank-based analysis

	Doripenem (N=374) n(%)	Levofloxacin (N=374) n(%)	DOOR probability (95% CI)	p-value
DOOR			**51.0% (47.6%, 54.3%)**	**0.5758**
DOOR Component				
Clinical Failure	81 (21.7%)	113 (30.2%)	54.3% (51.1%, 57.4%)	
Infectious Complications	23 (6.1%)	5 (1.3%)	47.6% (46.2%, 49.0%)	
SAEs	25 (6.7%)	14 (3.7%)	48.5% (46.9%, 50.1%)	
Death	1 (0.3%)	0 (0.0%)	49.9% (49.5%, 50.2%)	
DOOR			**51.0% (47.6%, 54.3%)**	**0.5758**
Sequential Dichotomization of DOOR				
DOOR <= 1 - DOOR > 1	263 (70.3%)	253 (67.6%)	51.3% (48.0%, 54.6%)	
DOOR <= 2 - DOOR > 2	356 (95.2%)	364 (97.3%)	48.9% (47.5%, 50.3%)	
DOOR <= 3 - DOOR > 3	372 (99.5%)	373 (99.7%)	49.9% (49.3%, 50.4%)	
DOOR <= 4 - DOOR > 4	373 (99.7%)	374 (100.0%)	49.9% (49.5%, 50.2%)	



Howard-Anderson J et al. Clin Infect Dis 2023 76:1157-1165.

Applying the DOOR in a Clinical Trial

DORI-05: Grade-based analysis- Partial credit analysis summary

DOOR (Most desirable to least desirable)	Grading key 1	Grading key 2	Grading key 3	Grading key 4	Grading key 5
Alive with no events	100	100	100	100	100
Alive with 1 event	100	100	100	0	80
Alive with 2 events	100	100	0	0	60
Alive with 3 events	100	0	0	0	40
Death	0	0	0	0	0

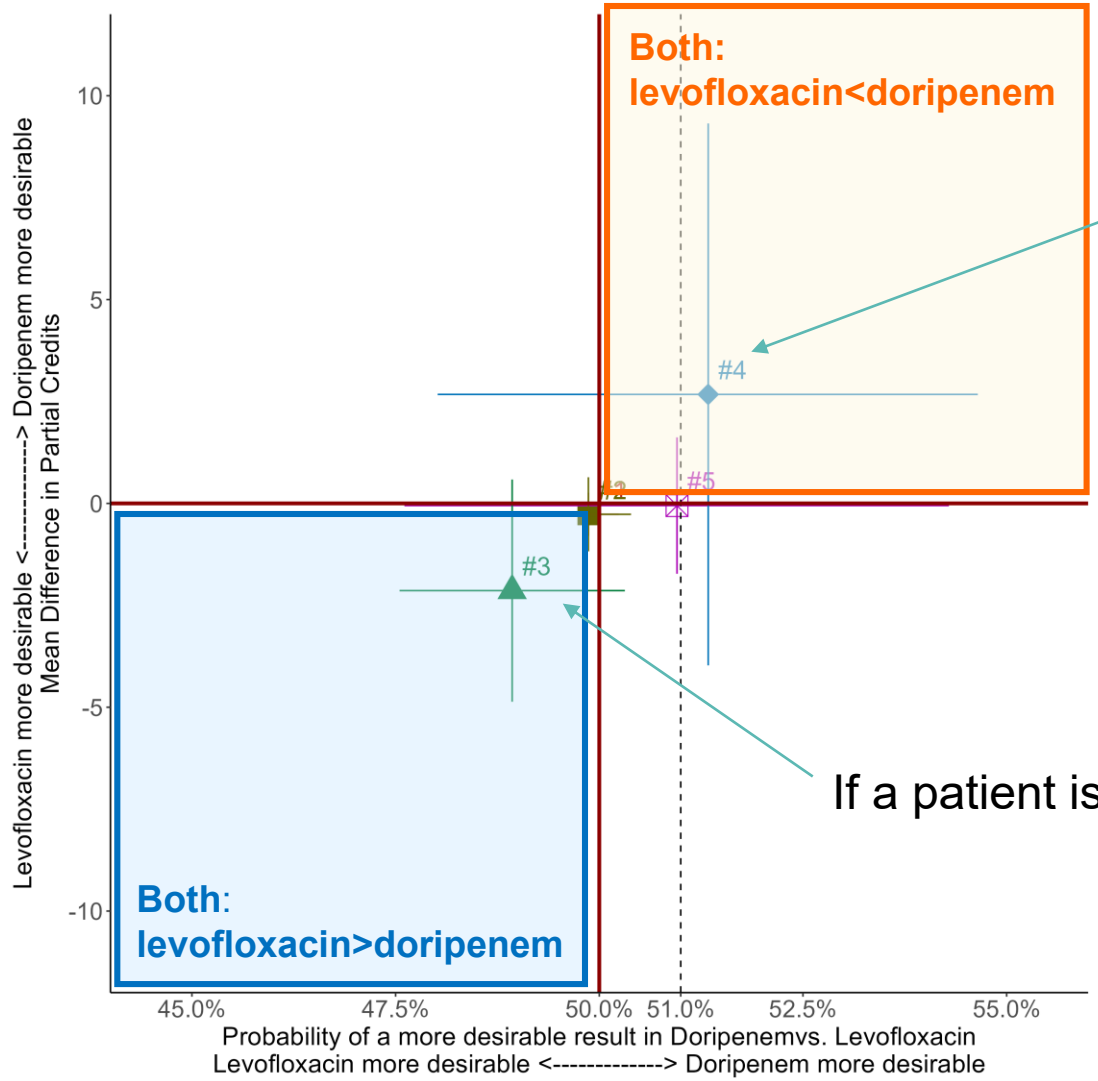
Statistics	DOR	LEV	DOR	LEV	DOR	LEV	DOR	LEV	DOR	LEV
Mean (SD)	99.7(5.2)	100.0(0.0)	99.5 (7.3)	99.7 (5.2)	95.2 (21.4)	97.3 (16.2)	70.3 (45.7)	67.6 (46.8)	92.9 (12.4)	92.9 (10.8)
Diff. in means(95%CI)	-0.3 (-0.8, 0.3)		-0.2 (-1.2, 0.6)		-2.1 (-4.9 , 0.6)		2.7 (-4.0 , 9.3)		0.0 (-1.7 , 1.6)	
P-value	0.3180		0.5635		0.1237		0.4299		0.9500	
DOOR probability (%) (95%CI)	49.9 (49.5 , 50.2)		49.9 (49.3, 50.4)		48.9 (47.5, 50.3)		51.3 (48.0, 54.6)		51.0 (47.6, 54.3)	
P-value	0.3173		0.5632		0.1236		0.4296		0.5758	

DOR: Doripenem; LEV: Levofloxacin

Howard-Anderson J et al. Clin Infect Dis 2023 76:1157-1165.

Applying the DOOR in a Clinical Trial

DORI-05: Grade-based analysis- Difference in partial credits vs DOOR probability



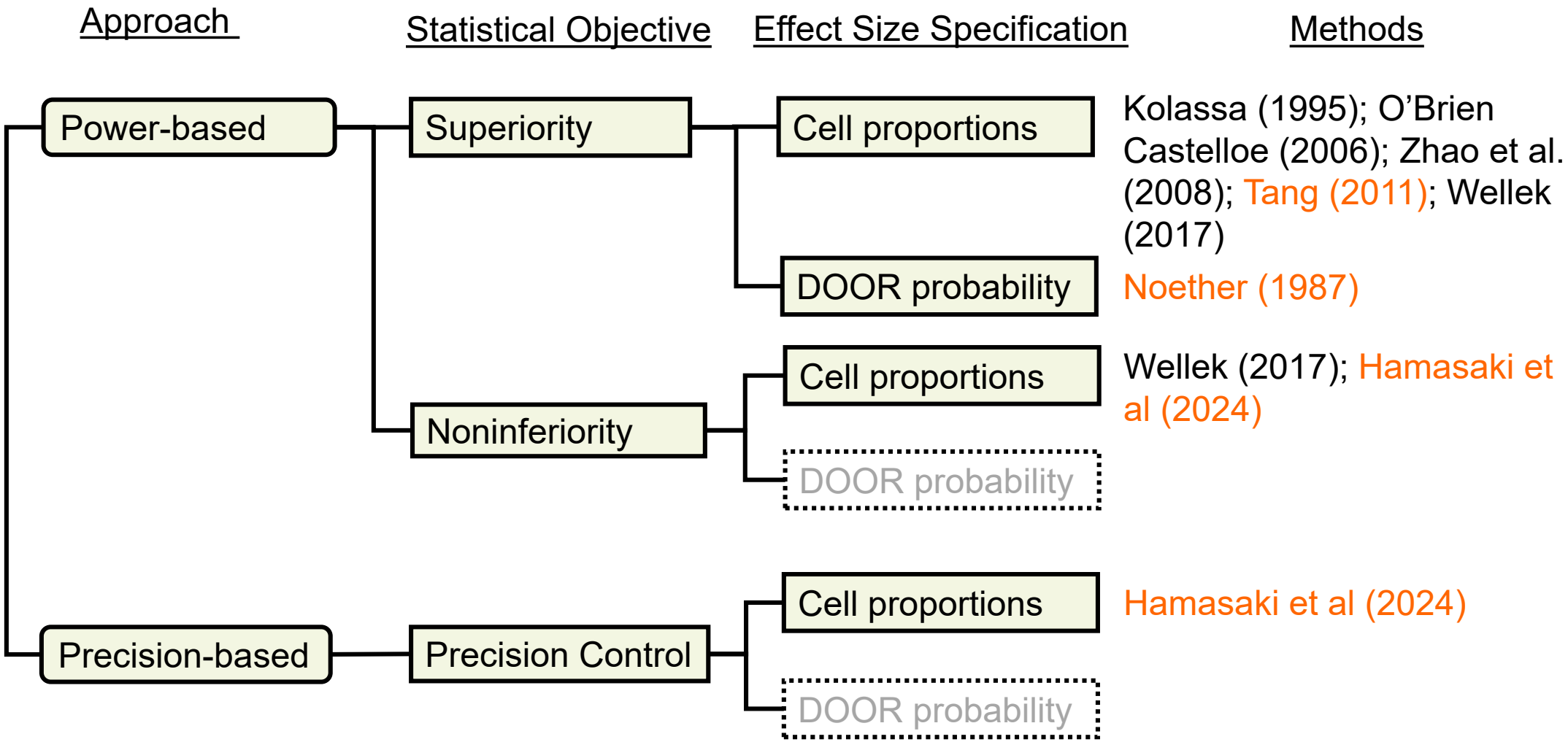
If a patient prefers **"No events"**

If a patient is OK with "**One event**"

Howard-Anderson J et al. Clin Infect Dis 2023 76:1157-1165.

Designing Clinical Trials with DOOR

Sample size determination methods based on the DOOR probability- available in our app



Noether GE. J Am Stat Assoc 1987; 82:645–647. Kolassa JE. Stat Med 1995; 14:1577–1581. Zhao YD et al. Stat Med 2008; 27:462–468. Tang Y. Stat Med 2011; 30:3461-3470. Tang Y Commun. Stat.-Simul. Comput 2016; 45:240-251,Wellek S. Stat Med 2017; 36:799-812. Hamasaki T et al. Biostatistics in Biopharmaceutical Research & Development: Clinical Trial Design, 137-159

Designing Clinical Trials with DOOR

Challenges in sample size determination– Effect size specification

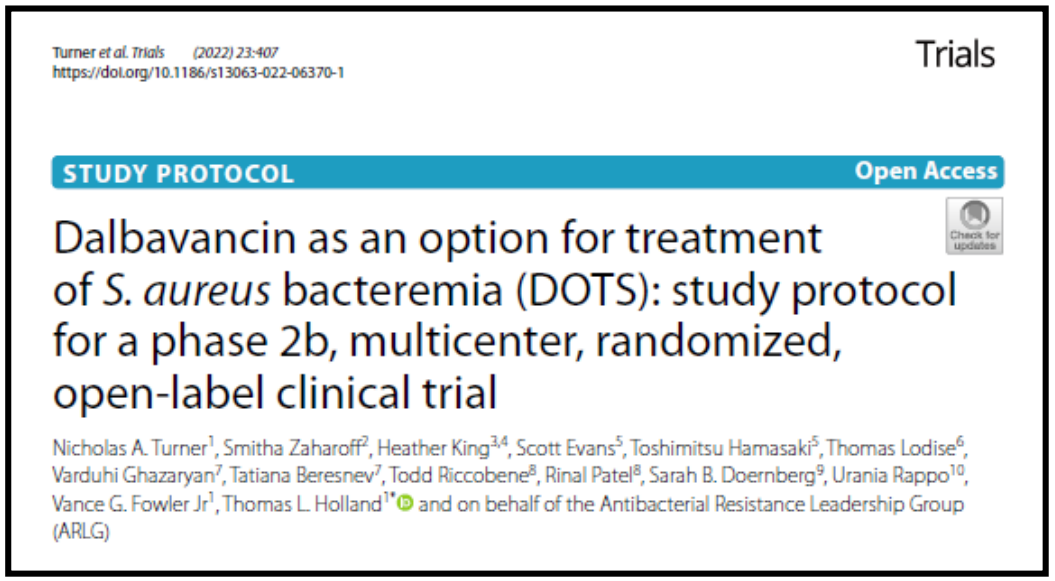
DOOR probability ⟺ Cell proportions

- Less parameters need to be specified (only one parameter- the DOOR probability to be detected)
- A larger sample size (generally 20% larger)

- More parameters need to be specified ($(2K - 2)$ cell proportions), but no or limited data is available
- A smaller sample size

Designing Clinical Trials with DOOR

An example



Turner et al. Trials (2022) 23:407
https://doi.org/10.1186/s13063-022-06370-1

Trials

STUDY PROTOCOL Open Access

Dalbavancin as an option for treatment of *S. aureus* bacteremia (DOTS): study protocol for a phase 2b, multicenter, randomized, open-label clinical trial

Nicholas A. Turner[1], Smitha Zaharoff[2], Heather King[3,4], Scott Evans[5], Toshimitsu Hamasaki[5], Thomas Lodise[6], Varduhi Ghazaryan[7], Tatiana Beresnev[7], Todd Riccobene[8], Rinal Patel[8], Sarah B. Doernberg[9], Urania Rappo[10], Vance G. Fowler Jr[1], Thomas L. Holland[1*] and on behalf of the Antibacterial Resistance Leadership Group (ARLG)

- A phase 2 randomized controlled trial
- Implemented the DOOR methodology
  - ❑ 5 level DOOR outcome including QoL as a tie-breaker

Plan

- Used Noether (1987), which requires the DOOR probability to be detected
  - ❑ No data was available

Analyses

- Planed to estimate the variance of the DOOR probability based on the cell proportions
  - ❑ The sample size would provide greater power to evaluate the outcome at the final analysis

Designing Clinical Trials with DOOR

Sensitivity assessment

- Important to investigate the sensitivity of the sample size to a variety of deviations from the assumptions
  - Power given a range of the DOOR probability (DOOR outcome distributions)?
  - Power given a range of sample size?

- Our developed online sample size assessment tool
  - Simulation function
  - Graphical tool to illustrate the behavior of power, given a range of the sample size and the DOOR probability (no commercial software does not have this tool!)

Designing Clinical Trials with DOOR

Online sample size assessment tool

DOOR: Power and Sample Size Assessment

Assessment

Approach
- (•) Power
- ( ) Precision

Type of Comparison
- (•) Superiority
- ( ) Non-inferiority

Solve for
- (•) Sample Size
- ( ) Power

DOOR Probability to Be Detected

DOOR Probability [Test >= Control] Defined by
- (•) DOOR Category Proportions (%)
- ( ) DOOR Probability (%)

No. of DOOR Categories (Maximum: 10)

5

DOOR Category Proportions (%) by Intervent
(Rank 1: most desirable to Rank 5: least desi

	Test	Control
Rank 1		
Rank 2		
Rank 3		
Rank 4		
Rank 5		
Total (%)	0	0

Calculated DOOR Probability: NA (%)

Configurations/Settings

One or Two-sided Test
- ( ) One-sided
- (•) Two-sided

Significance Level (α)
(e.g., 0.05, 0.025)

0.05

Allocation Ratio
(e.g., 0.5 means equally sized group)

0.5

Desired Power (1-β) (%)
(e.g., 80, 90)

80

DOOR Probability of Null Hypothesis (%)

50

Method
- (•) Method by Tang (2011)
- ( ) Normal Approximation
- ( ) Method by Noether (1987)

Assessment by Simulation

Power Evaluation by Simulation
- (•) No
- ( ) Yes

The DOOR Methodology

Summary

- Place interest on pragmatic questions to match their clinical importance
  - Implies a patient-centric benefit:risk focus
- The DOOR
  - Patient-centric paradigm for the design, data monitoring, analysis, interpretation, and reporting of clinical trials and other studies based on benefit-risk evaluation
  - Uses outcomes to analyze patients for a closer reflection of the effects on patients
  - Robust analyses
  - Online tools for the DOOR methodology
  - Ongoing work
    - Covariate-adjusted analysis / stratified analysis
    - Interim monitoring (group sequential and adaptive designs)
    - Subgroup evaluation
    - Meta-analyses
    - Longitudinal, time-to-event type DOOR outcomes

Milken Institute School of Public Health
THE GEORGE WASHINGTON UNIVERSITY

THE BIOSTATISTICS CENTER

SCT
SOCIETY FOR CLINICAL TRIALS

The DOOR is Open!

THANK YOU

https://methods.bsc.gwu.edu/

Duke Clinical Research Institute

Discussion, Perspectives and Experience using Hierarchical Endpoints in clinical trials

Frank W. Rockhold, PhD
Professor Biostatistics & Bioinformatics
Duke University Medical School
Society for Clinical Trials 2025, Vancouver BC

Disclosure Statement –
Frank W Rockhold, PhD

Research Funding: NIH, PCORI, DCRI, FDA, Astra Zeneca, American Regent, Alzheimer's Drug Discovery Foundation, Gates Foundation, BMS, Bayer, Pfizer, Priovant

IDMCs: Merck, AstraZeneca, Lilly, Novartis, Sanofi, UCB, BMS, Amgen, Biogen, BridgeBio, AskBio, Amgen, Cook, Pulmocide, Reunion, Priovant

Consulting: Clover/CEPI, Inventprise (Gates), Mountainfield

Boards: European Medicines Agency Technical Advisory Group, Frontier Science Foundation, California Institute for Regenerative Medicine, Doctor Evidence Medical Strategy, Clover Scientific Advisory Board

Equity Interest: GlaxoSmithKline, DataVant, Spencer, Doctor Evidence, Clover, Mountainfield

Comments on Hierarchal Endpoints

- Hierarchal Endpoint (HEP) approaches prioritize important events
  - Patients' worst outcome is always counted
  - Improves on weakness of composite endpoint assumptions
  - F-S statistic innovative but did not give a usable estimate of effect
- Win Ratio a measure developed to better interpret the F-S statistics and primarily employed in CV (HF) and Renal trials
- A useful measure of “net clinical benefit”, Generalized Pairwise Comparisons, Developed initially for oncology trials
- DOOR is a method to accommodate and summarize benefits and risks within a patient. Developed originally infectious disease trials.

Statistical Approaches for the WR

- Win ratio ignores ties so some use:
  - Win odds = (# wins + ½ # ties) / (# losses + ½ # ties)
  - Slightly smaller but bigger impact with more ties
  - Useful alternative, or just confusing?
- Win difference =% wins -% losses cumulative across the hierarchy, a complement to the WR

Win Ratio

- Is a way to interpret a HEP, it is not an “endpoint” in itself
- Can incorporate different endpoints (time to event, binary, count, ordinal, continuous, etc.).
- Does not require assumptions like proportional hazards for time to event endpoint.
- Does assume a common observation time
- WR magnitude can be hard to know ahead of time for the purposes of study design

Win Ratio Challenges

- Covariates and Stratification are tricky (research in progress)
- Interim analysis and data monitoring
- Clinical interpretation especially with continuous hierarchal endpoints.
- WR does not always “win” in terms of power
  - Choose the design with the clinical relevance, interpretability, and patient focus and then maximize power
  - Adding an endpoint for the purpose of “breaking ties” may not always improve power or help with clinical interpretation
- Non-inferiority- work in progress

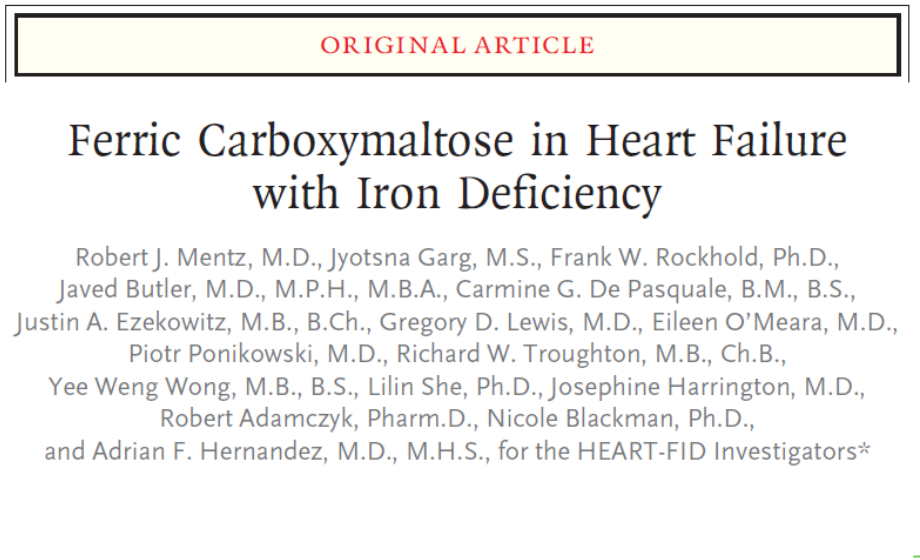
Common misunderstandings (Pocock 2023)

- Win ratio=1.36 does not mean that patients are 36% more likely to benefit on Empagliflozin than on placebo, though it is not seriously misleading.
- Does not mean that 36% of patients benefitted from Empagliflozin.
- The precise meaning is that of all patient pairs for which there was a preference (i.e. ignoring ties) there were 36% more wins on treatment than on placebo.
- Alternatively, win ratio=1.36 means that for any untied pair of patients, the odds that the winner is Empagliflozin (rather than placebo) is 1.36.
- Also, the win ratio is not the inverse of the HR except under certain circumstances, although again not seriously misleading

Examples

- HEART-FID
- ATTRIBUTE-CM

HEART-FID Trial



The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Ferric Carboxymaltose in Heart Failure with Iron Deficiency

Robert J. Mentz, M.D., Jyotsna Garg, M.S., Frank W. Rockhold, Ph.D., Javed Butler, M.D., M.P.H., M.B.A., Carmine G. De Pasquale, B.M., B.S., Justin A. Ezekowitz, M.B., B.Ch., Gregory D. Lewis, M.D., Eileen O'Meara, M.D., Piotr Ponikowski, M.D., Richard W. Troughton, M.B., Ch.B., Yee Weng Wong, M.B., B.S., Lilin She, Ph.D., Josephine Harrington, M.D., Robert Adamczyk, Pharm.D., Nicole Blackman, Ph.D., and Adrian F. Hernandez, M.D., M.H.S., for the HEART-FID Investigators*

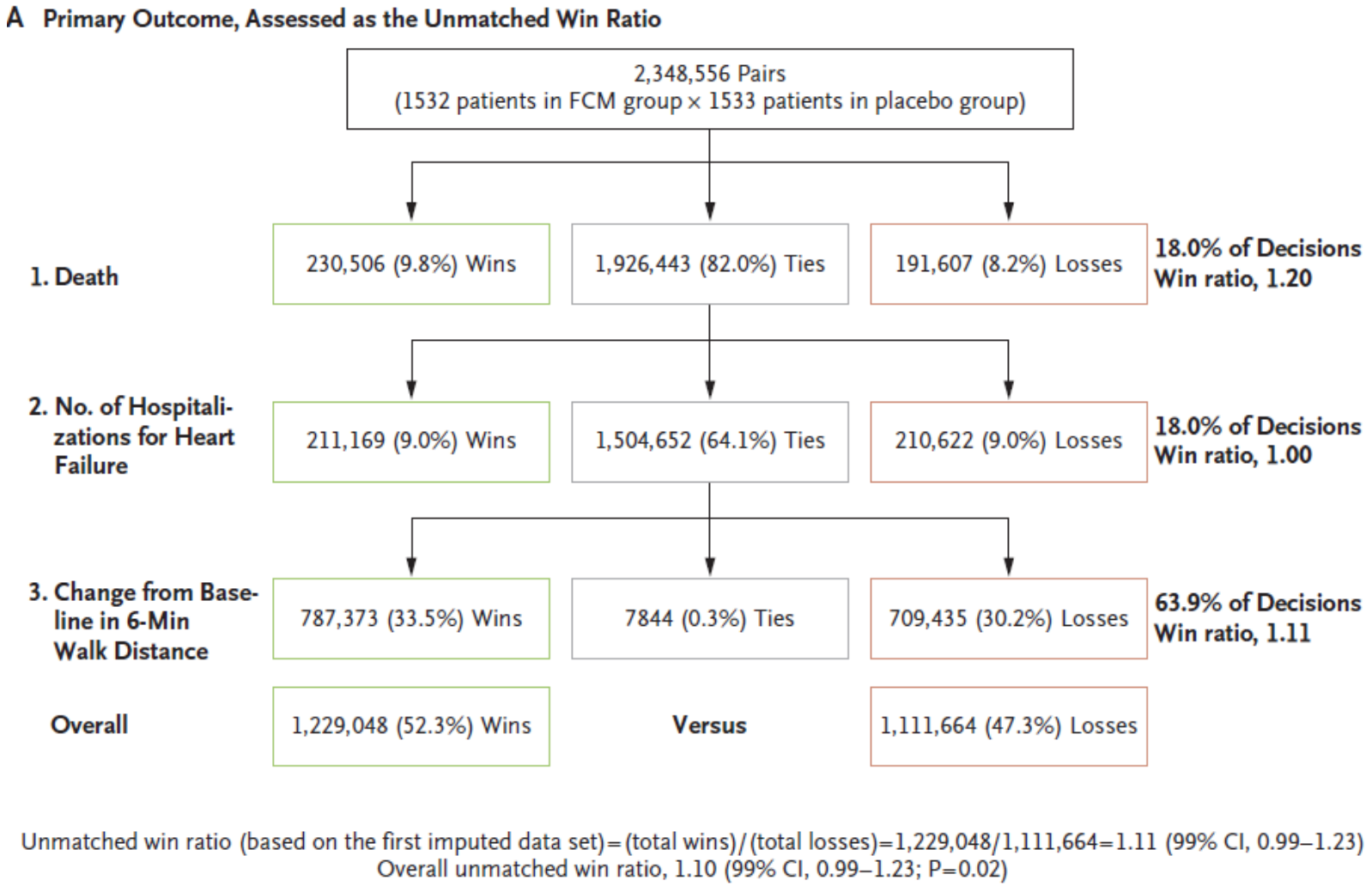
Hierarchical Endpoints

1. Time to death by 1 year.
2. Number of heart failure hospitalizations by 1 year.
3. Change of 6-minute walk distance from baseline to 6 months.

Heart-FID- (designed 2016)

- FDA wanted adequate clinical endpoint given it was new and worked well-
  - Secondary EP of Death/CV Hosp) was also a key focus.
- Number of events exceeded projected in 12 months
- Pandemic slowed Recruitment so avg time on study exceeded 12 mons
- Well powered did not yield the predicted treatment effect
- Interim analyses are challenging for HEP's so for this trial based on a was based on secondary composite endpoint

Example of Win Ratio: HEART-FID trial



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)

Example of Win Ratio: HEART-FID trial

- Double-blind, randomized (1:1) trial of patients with heart failure, left ventricular ejection fraction of 40% or less, and iron deficiency
- Treatments: intravenous ferric carboxymaltose or placebo, in addition to standard therapy for heart failure.
- Primary outcome: hierarchical composite of 1) death within 12 months after randomization, 2) hospitalizations for heart failure within 12 months after randomization, or 3) change from baseline to 6 months in the 6-minute walk distance.
- Primary analysis results: unmatched WR = 1.10, 99% CI (0.99-1.23), p=0.02 (Wilcoxon-Man-Whitney test)

ATTRIBUTE-CM Trial (Acoramidis)

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Efficacy and Safety of Acoramidis in Transthyretin Amyloid Cardiomyopathy

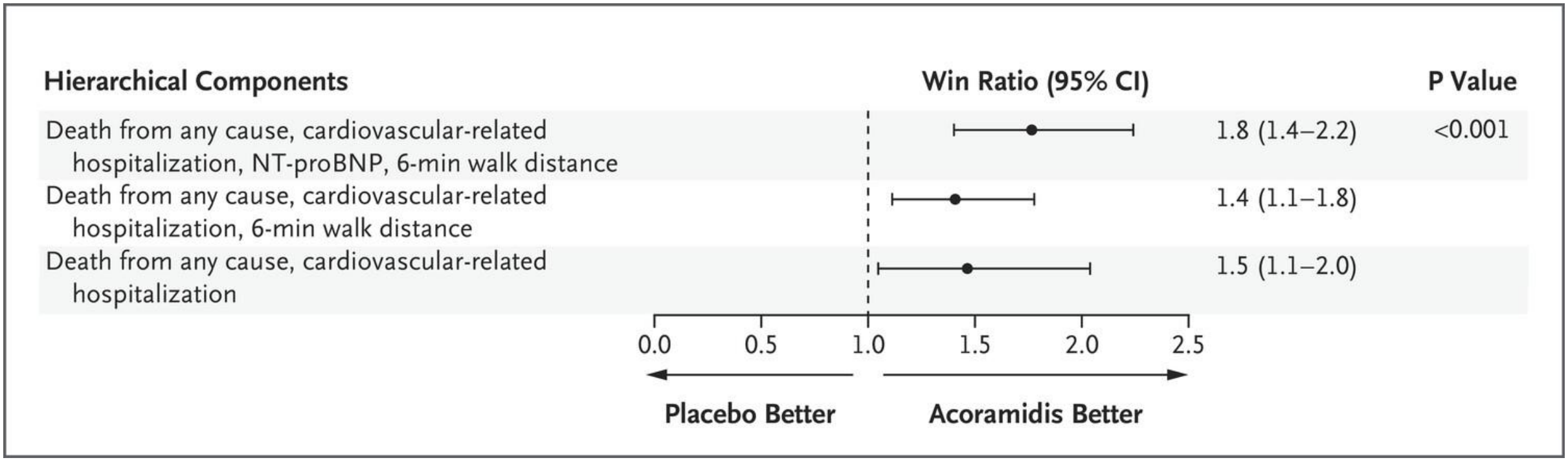
J.D. Gillmore, D.P. Judge, F. Cappelli, M. Fontana, P. Garcia-Pavia, S. Gibbs, M. Grogan, M. Hanna, J. Hoffman, A. Masri, M.S. Maurer, J. Nativi-Nicolau, L. Obici, S.H. Poulsen, F. Rockhold, K.B. Shah, P. Soman, J. Garg, K. Chiswell, H. Xu, X. Cao, T. Lystig, U. Sinha, and J.C. Fox, for the ATTRibute-CM Investigators*

Hierarchical Endpoints

1. **Time to all cause mortality**
2. **Number of CV hospitalizations**
3. **(added later) Change in NT-proBNP from baseline to month 30 (threshold of 500 pg/mL to win)**
4. **Change in 6-minute walk distance from baseline to month 30**

Acoramidis WR Table

Acoramidis Results Sensitivity



Generalized Pairwise Comparisons (GPC)

- GPC compares each pair of patients, one from the treatment group and one from the control group, and assigns a score based on which patient had a more desirable outcome.
- It can analyze a large set of outcomes, including various aspects of the disease evolution and patient quality of life.
- GPC is based on raw measurements instead of a single summary measure, which allows for a more detailed analysis.
- The treatment effect in GPC can be expressed as the net treatment benefit, the success odds, or the win ratio.
- GPC is particularly useful in situations where a single composite endpoint doesn't fully capture the clinical benefit of a treatment.

Desirability of Outcome Ranking (DOOR)

- Desirability of Outcome Ranking (DOOR):
- DOOR classifies patients into ordinal categories based on overall clinical outcome, considering both benefits and harms.
- The number and definition of categories are tailored to the specific clinical disease.
- DOOR is used in infectious disease trials, where it allows for a more nuanced assessment of treatment effectiveness.
- Useful Summary in Data Monitoring Benefit Risk even without inference

Three Methods to evaluate Hierarchal Endpoints

- All are improvements over composite endpoints
- All are useful and have different advantages and disadvantages
- All have assumptions
- All have challenges at the study design phase in terms of expressing the magnitude of effect.
- W-R can be adapted for benefit risk, but GPC and DOOR are particularly useful at examining BR at group and individual level.