

Interpreting the Win Ratio: Five Key Challenges for Clinicians

Based on: Thomsen et al. arXiv:2504.05909

Julie F. Furberg | Novo Nordisk A/S | DSTS Two Day Meeting 2026

04 May 2026

Presenting on behalf of original
presenter Henrik F. Thomsen
(slides prepared by him too 😊)

Disclaimer

The views and opinions expressed in this presentation are those of the author and do not necessarily reflect the official policy or position of Novo Nordisk A/S.

This presentation is based on two companion articles:

1. Thomsen et al. "Interpreting the Win Ratio: Five Key Challenges for Clinicians"

Clinical perspective article (submitted)

2. Thomsen et al. "Interpretational Challenges of the Win Ratio in Analyzing HCEs in CKD"

Statistical perspective article (arXiv:2504.05909)

Why Do We Run Clinical Trials?

Randomised controlled trials are the foundation of evidence-based medicine and regulatory decision-making

The Fundamental Problem

- **We want to know if a treatment causes** improvement, not just correlates with it
- Observational data cannot reliably separate treatment effects from confounding

The RCT Solution

- Randomisation balances known and unknown confounders, enabling valid causal inference
- FDA and EMA drug approvals are built on RCT evidence

The Estimand: The Question First

ICH E9(R1): five components define the question before any analysis begins:

- **Population:** who is in the trial?
- **Treatment:** what is the intervention (including adherence assumptions)?
- **Variable:** what outcome is measured?
- **Intercurrent events:** how are death, dropout etc. handled?
- **Population-level summary:** e.g., mean difference, win ratio

“The clinical question should drive the choice of analysis method – not statistical convenience.”

From Single Endpoints to Composite Endpoints

Real diseases involve multiple outcomes of varying severity; single endpoints often cannot capture the full picture

Single Endpoint

The problem

- Which outcome matters most?
Death, organ failure, functional decline – in reality, all of them
- Trials powered on one rare event require very large sample sizes

Traditional Composite

The partial solution

- Combines outcomes into one endpoint: e.g., MACE (CV death + MI + stroke)
- Increases power, but treats all components as equally important
- *Is a hospitalisation really equivalent to death?*

Hierarchical Composite

The richer approach

- Orders outcomes by clinical severity: most severe takes precedence
- Captures the full disease trajectory: from catastrophic events down to early signs
- Hierarchy set by clinical judgement, not statistical convenience

“The hierarchy is not a statistical choice: it is a clinical statement about what matters most.”

Introducing the Win Ratio

The Win Ratio is the natural analytical tool for hierarchical composite endpoints – and it is now regulatory standard

The Pairwise Comparison

1. Pair each treated patient with each control patient
2. Ask: **who did better?** Assess on the highest-priority component first
3. If tied on that component, move down the hierarchy to the next
4. Continue until a winner is found; declare a tie only if all components are tied
5. Repeat for all treatment–control pairs

The Win Ratio Statistic

$$\text{WR} = \frac{\text{Wins}(\text{treatment})}{\text{Wins}(\text{control})}$$

- **WR > 1** treatment favoured
- **WR < 1** control favoured

Clinically communicable: “Treated patients won more comparisons than control patients”

“The Win Ratio is powerful and increasingly used in regulatory settings – but does it always answer the question we think it does? Today we examine five key challenges.”

Five Key Interpretational Challenges

1

**What Question Does
the Win Ratio
Answer?**

Defining the estimand
and clinical question

2

**Population-Level vs.
Individual-Level
Effects**

Hand's Paradox:
opposite conclusions
are possible

3

**Non-Transitivity of
Treatment Rankings**

$A > B$ and $B > C$ does
not imply $A > C$

4

**Dependence on
Outcome Variability**

Same effect, different
WR based on study
design

5

**Non-Collapsibility
Across Populations**

Overall WR may differ
from subgroup-specific
WRs

Background: HCEs and the Win Ratio

Hierarchical Composite Endpoints

Multiple outcomes ordered by severity:

1. All-cause mortality
2. Kidney replacement therapy
3. Sustained GFR <15 mL/min/1.73m²
- 4-6. Sustained GFR declines (≥57%, 50%, 40%)
7. GFR slope

The Win Ratio Method

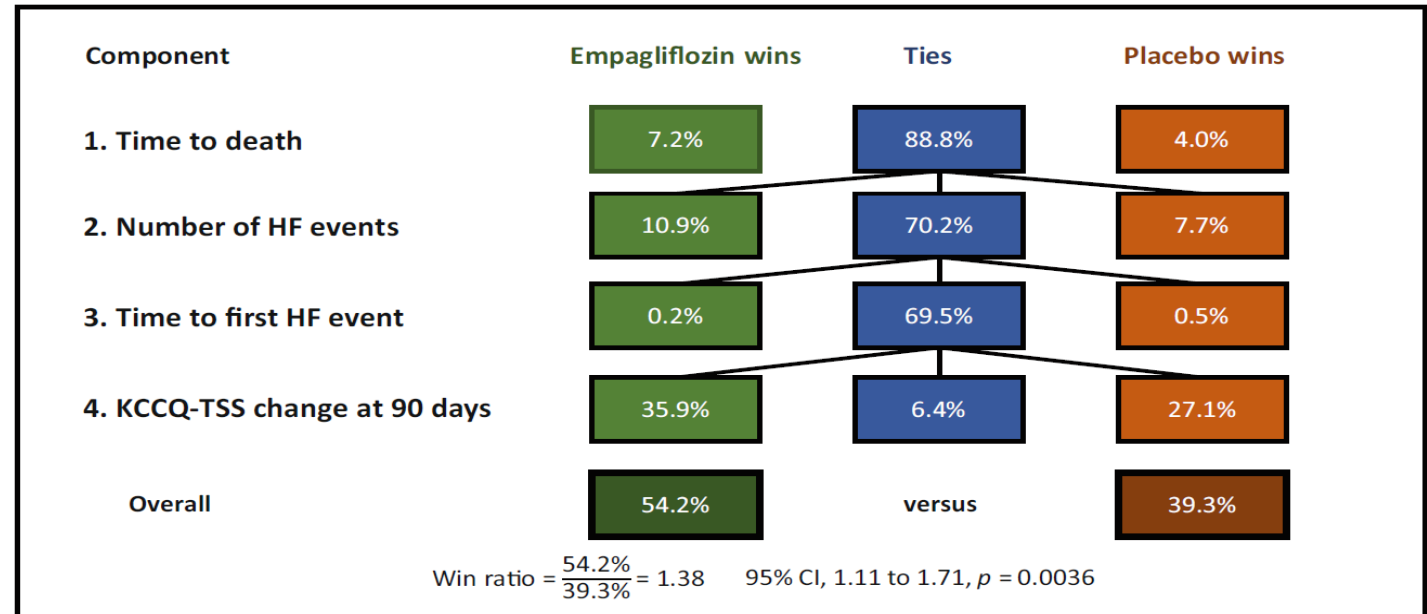
Compare each treatment patient to each control:

- Determine winner on highest-priority component
- If tied, move to next component
- Continue until winner found or all tied

$$WR = P(\text{treatment wins}) / P(\text{control wins})$$

WR > 1 favours treatment |

WR < 1 favours control



EMPULSE trial (heart failure) - Pocock et al., Eur J Heart Fail, 2023

Advantages: potential increased statistical power | mitigates competing risks |
no multiplicity adjustment | captures full disease trajectory

FDA approvals: Tafamidis & Acoramidis (cardiac amyloidosis)

CHALLENGE 1

What Question Does the Win Ratio Answer?

The WR answers "Is a treated patient more likely to do better than a control patient?" rather than "Will this patient benefit?" The question posed must be coherent: the HCE components should share a meaningful clinical relationship

Scenario 1a**Surrogate for clinical events**

Continuous endpoint serves as early surrogate (e.g., GFR decline as surrogate for ESRD)

→ HCE captures both subclinical progression and clinical outcome

✓ Coherent clinical question

Scenario 1b**Independent outcome of interest**

Continuous endpoint captures a distinct concept (e.g., KCCQ score in HF trial)

→ HCE conflates mortality effects with symptom effects

✗ May not address clinical question

Scenario 2**Holistic disease assessment**

Multiple distinct but related outcomes (death, dialysis, GFR slope)

→ Captures multifaceted nature of disease progression

✓ Appropriate if holistic view is goal

CHALLENGE 2

The WR as a Population-Level Summary

Population-level WR

Compares all subjects in the treatment arm to all subjects in the control arm

$$\text{WR} = \text{P}(\text{treated wins}) / \text{P}(\text{control wins})$$

Individual-level WR

Compares each subject to themselves: what would their outcome have been under the alternative treatment?

Not identifiable in a randomized trial!

These two measures are NOT guaranteed to agree. They can even point in opposite directions.

The mean difference is both an individual- and population-level measure. The WR is purely population-level.

Key distinction

The WR answers a population-level question. It does NOT directly tell us:

"If I take this treatment, am I more likely to do better?"

This individual-level question is what patients and clinicians actually need answered

Clinical implications

- Hand's Paradox: population- and individual-level measures can yield **opposite conclusions**
- Unlike the mean difference (both individual- and population-level), the WR is purely population-level
- Matching approaches (Even & Josse, 2025) may approximate individual-level effects, though this remains under development

CHALLENGE 2

Hand's Paradox: Same Data, Opposite Conclusions

Patient	Y(1)	Y(0)
1	1	6
2	3	2
3	5	4

Higher values = better outcomes

Population-level WR

Compare all Y(1) vs. all Y(0):

Treatment wins: $3 > 2$ $5 > 2$ $5 > 4 = 3$

Control wins: $6 > 1$ $6 > 3$ $6 > 5$ $2 > 1$ $4 > 1$ $4 > 3 = 6$

WR = $3/6 = 0.5$ → Treatment appears harmful

Individual-level WR

Compare each patient to themselves:

Patient 1: $Y(1)=1 < Y(0)=6$ → control wins

Patient 2: $Y(1)=3 > Y(0)=2$ → treatment wins

Patient 3: $Y(1)=5 > Y(0)=4$ → treatment wins

WR = $2/1 = 2.0$ → Treatment appears beneficial

Mean difference: consistent either way

Individual-level: $\text{mean}(Y(0)-Y(1)) = [(6-1)+(2-3)+(4-5)]/3 = 1.0$ favouring control

Population-level: $\text{mean}[Y(0)] - \text{mean}[Y(1)] = 4.0 - 3.0 = 1.0$ favouring control

Mean difference gives the same answer either way. The Win Ratio does not.

The individual-level WR is not identifiable in a randomized trial because we never observe both potential outcomes for the same patient

CHALLENGE 3

Non-Transitivity of Treatment Rankings

Patient	A	B	C
1	2	1	3
2	4	6	5
3	9	8	7
Mean	5	5	5

Higher values = better outcomes (e.g.,
kg weight loss)

Win Ratio comparisons:

$WR(A \text{ vs } B) = 5/4 = 1.25$ A beats B

$WR(B \text{ vs } C) = 5/4 = 1.25$ B beats C

$WR(C \text{ vs } A) = 5/4 = 1.25$ C beats A!

A > B > C > A (circular!)

Mean difference comparison:

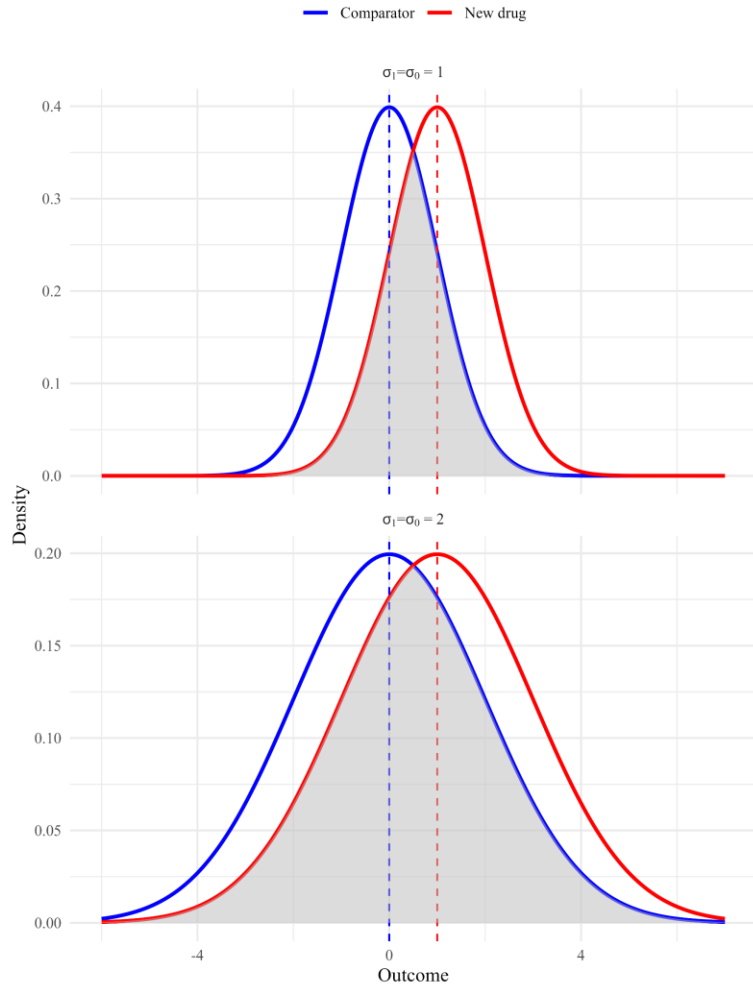
Mean A = Mean B = Mean C = 5 kg

All pairwise mean differences = 0. Correctly identifies equivalent efficacy.

Implication: indirect comparisons and network meta-analyses based on WRs cannot assume transitivity. This undermines their use for consistent treatment ranking.

CHALLENGE 4

The Win Ratio Depends on Variability



$\sigma = 1$ (low variability)

$$WR \approx 3.17$$

Minimal overlap — groups clearly separable

$\sigma = 2$ (high variability)

$$WR \approx 1.76$$

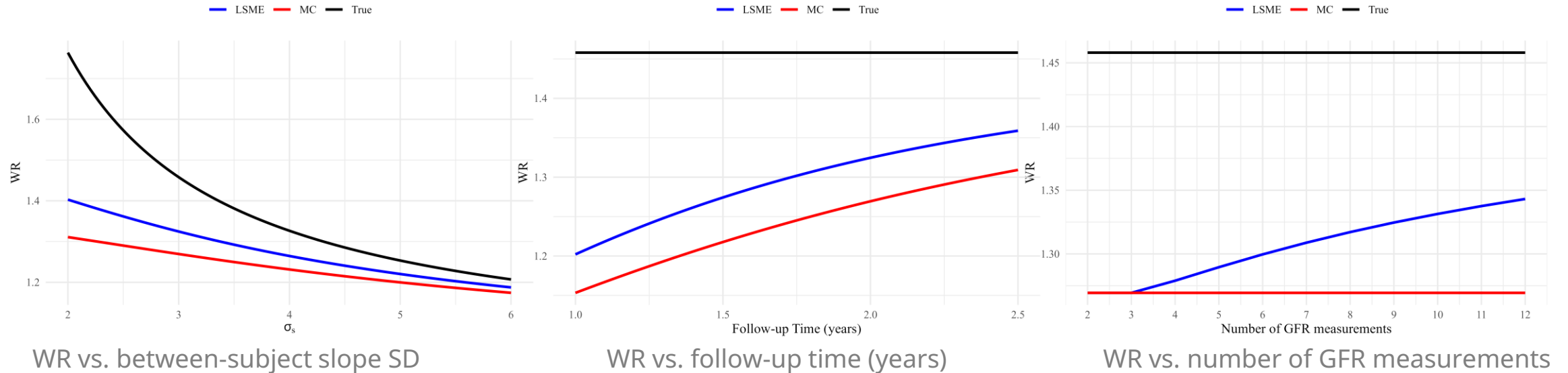
Substantial overlap — harder to discriminate

Same 1-unit mean difference in both cases!

The WR measures discrimination (Cohen's d),
not magnitude.

CHALLENGE 4

Study Design Affects Win Ratio Magnitude



CKD trial: GFR slope as HCE component

WR decreases as between-patient variability increases (left panel), even when the treatment effect is unchanged.

Longer follow-up (centre panel) and more frequent GFR measurements (right panel) increase WR by reducing measurement error.

Implication: comparing WR magnitudes across studies with different designs is problematic.

CHALLENGE 5

Non-Collapsibility Across Populations

	% of Pop.	Treatment mean	Control mean	WR
Subgroup 1	50%	55	50	3.17
Subgroup 2	50%	65	60	3.17
Overall	100%	—	—	2.18

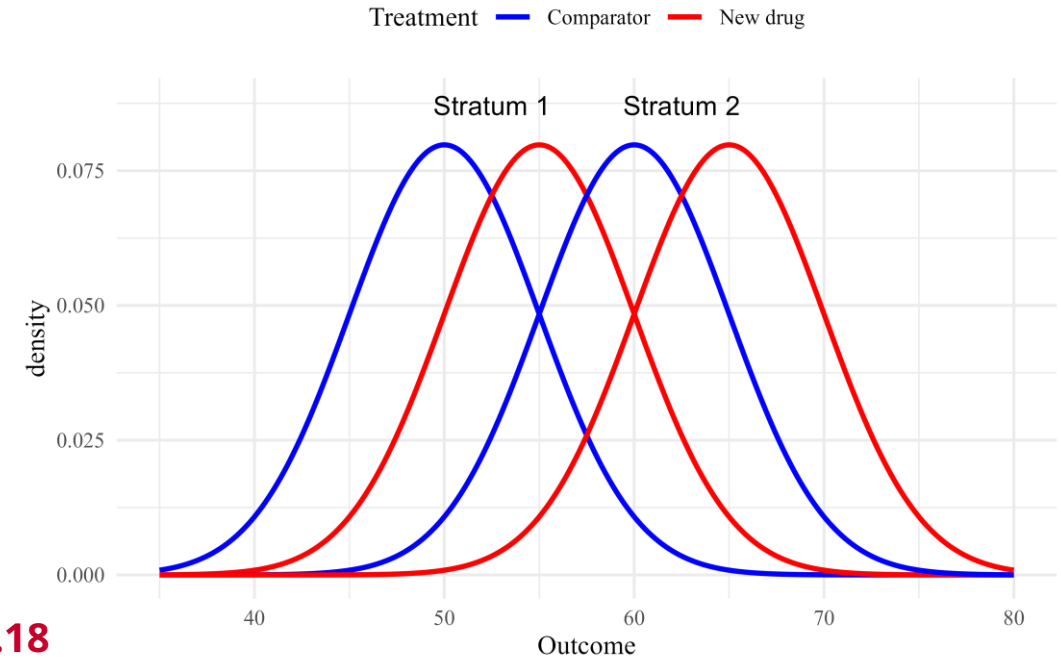
Normally distributed, common SD = 5. Treatment effect = 5 in both subgroups.

Identical within-subgroup WRs (3.17) yet overall WR is only 2.18

Cross-subgroup comparisons dilute the WR: a treatment patient from subgroup 1 (mean=55) vs. a control patient from subgroup 2 (mean=60) will often lose despite the treatment being effective.

Trials with identical treatment effects can report different overall WRs if subgroup proportions differ. This complicates meta-analyses and cross-trial comparisons.

Shared property with odds ratios and hazard ratios. A stratified WR yields the correct within-stratum estimate (3.17).



The Win Ratio in Context: Broader Considerations

The WR as a Discriminatory Measure

- Measures **stochastic superiority**: how often treatment beats control in pairwise comparisons, not the magnitude of benefit
- Analogous to non-parametric tests; should be reported alongside effect sizes rather than used as a standalone measure
- A WR of 1.5 means a treated patient wins 50% more pairwise comparisons. It does not mean each patient gains 50%.

Emerging Tools and Broader Reporting

- **Maraca plot**: visualises each HCE component's contribution to the overall WR, aiding clinical communication and transparency
- **Benefit-risk indices**: supplement the WR with absolute risk reduction and NNT for each component
- The WR is best used as **one tool in a broader reporting framework**, not as a standalone summary statistic

Practical Guidance for Clinicians

- Ask whether the WR reflects a clinically coherent question for this endpoint combination
- Dissect component contributions and do not rely on the WR alone to judge treatment benefit
- Exercise caution when comparing WRs across trials with different designs or follow-up durations

The clinical question should drive the choice of analysis method, not statistical convenience.

Practical Recommendations

- 1 Define the estimand first**

Choose the WR because it addresses the clinical question, not for statistical convenience. Pre-specify what the composite measures.
- 2 Examine individual components**

A WR driven by mortality reduction has different implications than one driven by GFR slope differences. Always report component analyses.
- 3 Recognize population-level interpretation**

The WR answers a population question, not an individual one. It provides evidence that treatment works, but does not directly quantify individual benefit.
- 4 Exercise caution with indirect comparisons**

Non-transitivity means that WR-based network meta-analyses cannot assume consistent treatment rankings.
- 5 Consider study design when comparing WRs**

WR magnitude reflects both treatment effect AND measurement characteristics. Trials with different designs may report different WRs for identical effects.
- 6 Interpret subgroup analyses carefully**

Differences between overall and subgroup-specific WRs may reflect non-collapsibility, not true effect modification. Seek formal interaction tests.

Thank You

Julie F. Furberg

Biostatistics, Novo Nordisk A/S

jukf@novonordisk.com

Co-authors: H. Thomsen, S. Gasparian, *J. Furberg*, C. Tasto, N. Rethemeier,
P. Schloemer, T. Wang, N. Jongs, Y. Du, T. Greene