

A series of overlapping, tilted rectangular outlines in the top-left corner of the slide, creating a complex geometric pattern.

Hierarchical Composite Endpoints: More Nuance, More Insight and ... More Confusion?

EFSPI Regulatory Statistics Workshop
11SEP2025

Henrik F. Thomsen & Mickaël De Backer



DISCLAIMER

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

All numerical examples are fictitious.

WHAT THIS TALK IS ABOUT

Two thin, dark lines intersect in the top right corner of the slide. One line is nearly horizontal, sloping slightly downwards from left to right. The other line is more vertical, sloping downwards from right to left.

WHAT THIS TALK IS ABOUT

Under the umbrella: hierarchical composite endpoints form a landscape

Two thin, dark lines intersect in the top right corner of the slide. One line is horizontal, and the other is diagonal, creating a simple geometric design.

WHAT THIS TALK IS ABOUT

Under the umbrella: hierarchical composite endpoints form a landscape

This is new, this is not new

WHAT THIS TALK IS ABOUT

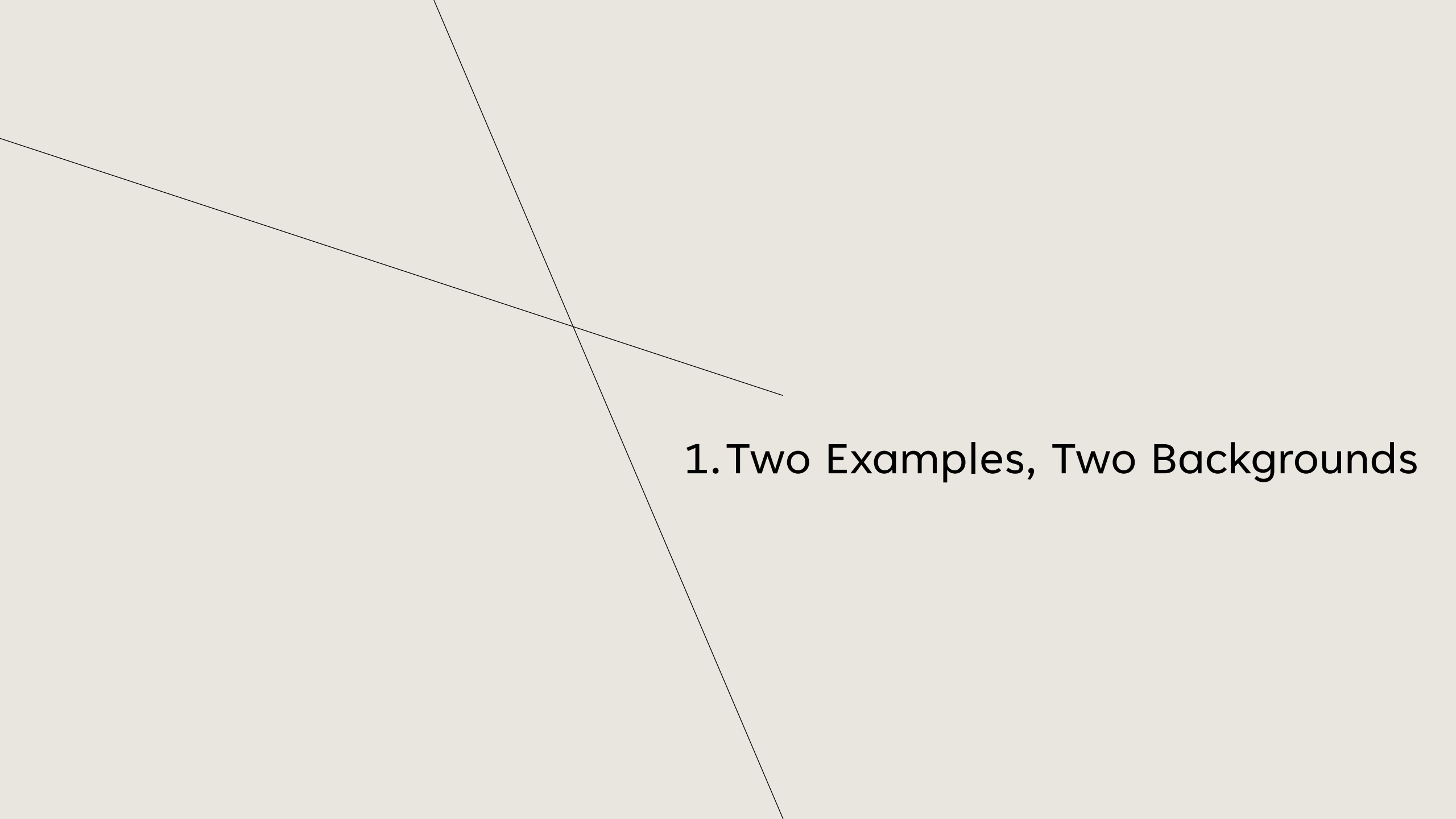
Under the umbrella: hierarchical composite endpoints form a landscape

This is new, this is not new

Multivariate: Everything Everywhere All At Once

OUTLINE

1. Two Examples, Two Backgrounds
2. A World Inside a World: Landscape of HCEs
3. Win Statistics: the Good, the Bad and the Misleading
4. The Baby and the Bathwater
5. Built-In Tensions

Two thin, dark gray lines intersect diagonally on a light gray background. One line runs from the top-left towards the bottom-right, and the other runs from the top-right towards the bottom-left. They cross each other in the upper-left quadrant of the image.

1. Two Examples, Two Backgrounds

FIRST EXAMPLE (PRE-ESTIMAND WORLD)

Clinical context

- Head & Neck Cancer
- Drug aimed at side-effect of radiotherapy: severe oral mucositis (SOM)
- Primary endpoint: incidence of grade 3 or 4 SOM (liquid diet only or alimentation not possible)

FIRST EXAMPLE (PRE-ESTIMAND WORLD)

Clinical context

- Head & Neck Cancer
- Drug aimed at side-effect of radiotherapy: severe oral mucositis (SOM)
- Primary endpoint: incidence of grade 3 or 4 SOM (liquid diet only or alimentation not possible)

Headline

“Primary endpoint of reduction in SOM incidence was not met in the trial”

FIRST EXAMPLE (PRE-ESTIMAND WORLD)

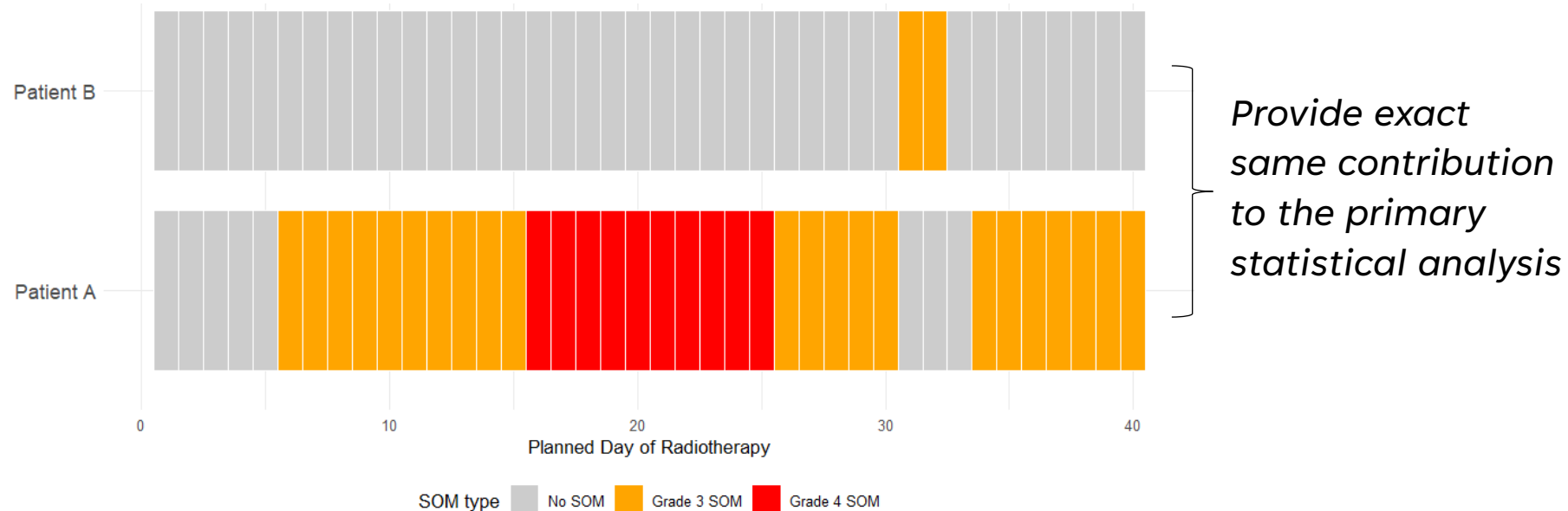
Clinical context

- Head & Neck Cancer
- Drug aimed at side-effect of radiotherapy: severe oral mucositis (SOM)
- Primary endpoint: incidence of grade 3 or 4 SOM (liquid diet only or alimentation not possible)

Headline

“Primary endpoint of reduction in SOM incidence was not met in the trial”

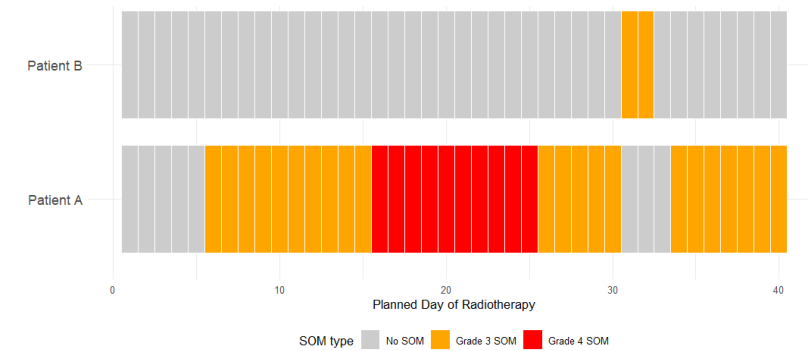
The (very simplified) journeys of two patients



FIRST EXAMPLE (PRE-ESTIMAND WORLD)

What are we saying?

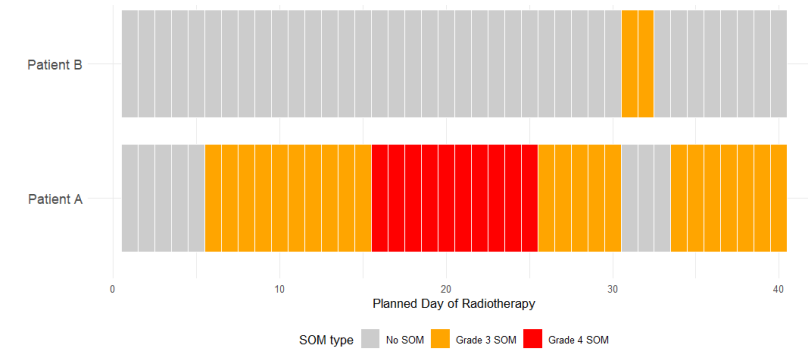
- Incidence is poor, statistically and clinically... *Not new*



FIRST EXAMPLE (PRE-ESTIMAND WORLD)

What are we saying?

- Incidence is poor, statistically and clinically... *Not new*
- For “efficacy”, incidence may not be enough:
 - i. Gr4 is worse than Gr3
 - ii. Longer SOM episodes are worse
 - iii. Earlier SOM episodes are worse



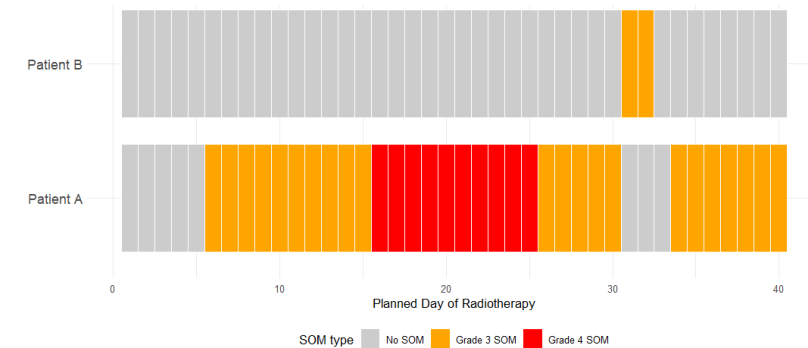
FIRST EXAMPLE (PRE-ESTIMAND WORLD)

What are we saying?

- Incidence is poor, statistically and clinically... *Not new*
- For “efficacy”, incidence may not be enough:
 - i. Gr4 is worse than Gr3
 - ii. Longer SOM episodes are worse
 - iii. Earlier SOM episodes are worse

In practice

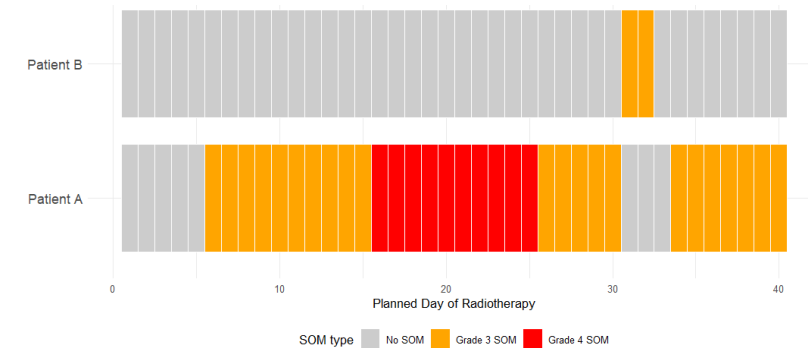
- We look at the different angles, but... separately



FIRST EXAMPLE (PRE-ESTIMAND WORLD)

What are we saying?

- Incidence is poor, statistically and clinically... *Not new*
- For “efficacy”, incidence may not be enough:
 - i. Gr4 is worse than Gr3
 - ii. Longer SOM episodes are worse
 - iii. Earlier SOM episodes are worse



In practice

- We look at the different angles, but... separately

Question

- Suppose:
 - Drug helps all aspects of SOMs,
 - Anticipated due to same mechanism of action,
 - Each signal separately is not strong enough (within reasonable RCTs)
- Wouldn't we still want to be able to detect an 'overall'* signal within a feasible trial?

SECOND EXAMPLE

Clinical context

- Stable coronary artery disease
- Intervention aimed at relieving angina episodes
- Anti-anginal medication allowed

SECOND EXAMPLE

Clinical context

- Stable coronary artery disease
- Intervention aimed at relieving angina episodes
- Anti-anginal medication allowed

Composite strategy in the estimand framework

Patient journeys are filled with ICEs

- Some change our understanding of what we measure (e.g., rescue medication)
- Some affect the existence of the outcome of interest (e.g., death)

SECOND EXAMPLE

Clinical context

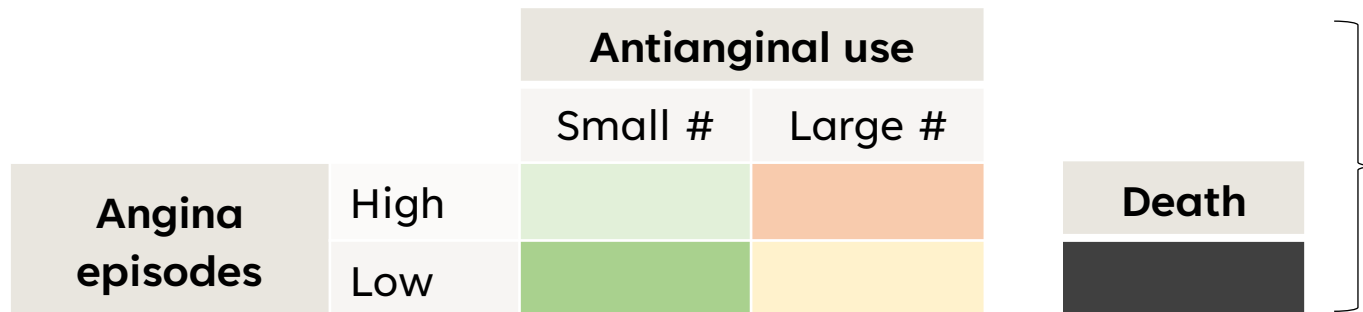
- Stable coronary artery disease
- Intervention aimed at relieving angina episodes
- Anti-anginal medication allowed

Composite strategy in the estimand framework

Patient journeys are filled with ICEs

- Some change our understanding of what we measure (e.g., rescue medication)
- Some affect the existence of the outcome of interest (e.g., death)

Idea (oversimplified – more later)



Generalizes, in spirit, existing practices:

- EDSS (Expanded Disability Status Scale) in Multiple Sclerosis
- mRS (modified Rankin scale) in cardiology

TWO EXAMPLES, TWO BACKGROUNDS

Example 1

- i. Gr4 is worse than Gr3
- ii. Longer SOM episodes are worse
- iii. Earlier SOM episodes are worse

↓ Top-down
incidence is
'not enough'

Ambition

Increase sensitivity to detect a signal*:

- Accumulate sources of signal
- No multiplicity adjustment

Similar spirit: time-to-first event analysis

TWO EXAMPLES, TWO BACKGROUNDS

Example 1

- i. Gr4 is worse than Gr3
- ii. Longer SOM episodes are worse
- iii. Earlier SOM episodes are worse

↓ Top-down
incidence is
'not enough'

Example 2

- i. Death
- ii. Large episode #, high rescue medication
- iii. Large episode #, low rescue medication
- iv. Small episode #, high rescue medication
- v. Small episode #, low rescue medication

↑ Bottom-up
initial interest
started from
Angina episodes

Ambition

Increase sensitivity to detect a signal*:

- Accumulate sources of signal
- No multiplicity adjustment

Similar spirit: time-to-first event analysis

Ambition

- Align with composite strategy in estimand framework (death)
- Multivariate aims to increase understanding:
 - Outcome value nuanced by rescue medication,
 - Joint analysis of benefits and risks

* Where there's vagueness, there's room for abuse

TWO EXAMPLES, TWO BACKGROUNDS

Example 1

- i. Gr4 is worse than Gr3
- ii. Longer SOM episodes are worse
- iii. Earlier SOM episodes are worse

↓
Top-down
incidence is
'not enough'

Example 2

- i. Death
- ii. Large episode #, high rescue medication
- iii. Large episode #, low rescue medication
- iv. Small episode #, high rescue medication
- v. Small episode #, low rescue medication

↑
Bottom-up
initial interest
started from
Angina episodes

Ambition

Increase sensitivity to detect a signal*:

- Accumulate sources of signal
- No multiplicity adjustment

Similar spirit: time-to-first event analysis

Ambition

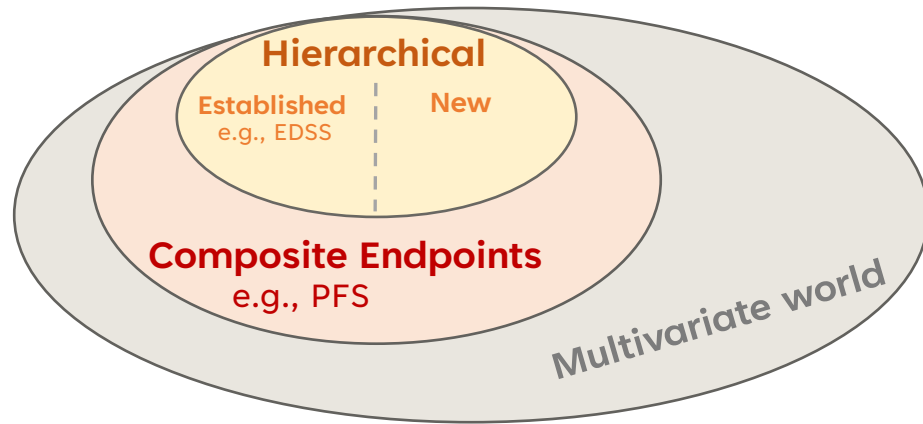
- Align with composite strategy in estimand framework (death)
- Multivariate aims to increase understanding:
 - Outcome value nuanced by rescue medication,
 - Joint analysis of benefits and risks

The umbrella: *some* order of desirability exists across what we measure

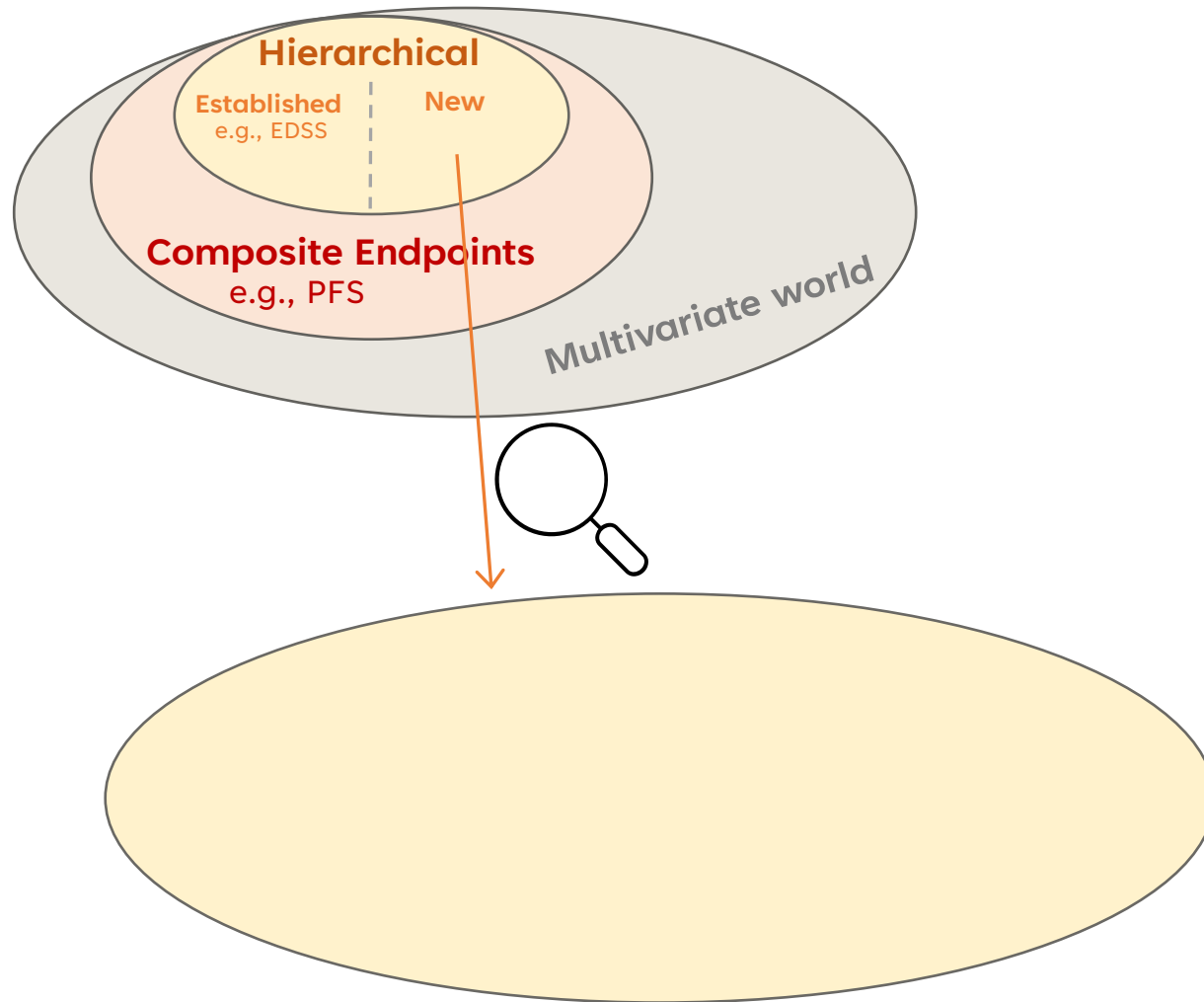
An abstract geometric design featuring two thin, dark grey lines that intersect on a light grey background. One line runs diagonally from the top-left towards the bottom-right, while the other runs from the top-right towards the bottom-left. The intersection point is located in the upper-left quadrant of the slide.

2. A World Inside a World: Landscape of HCEs

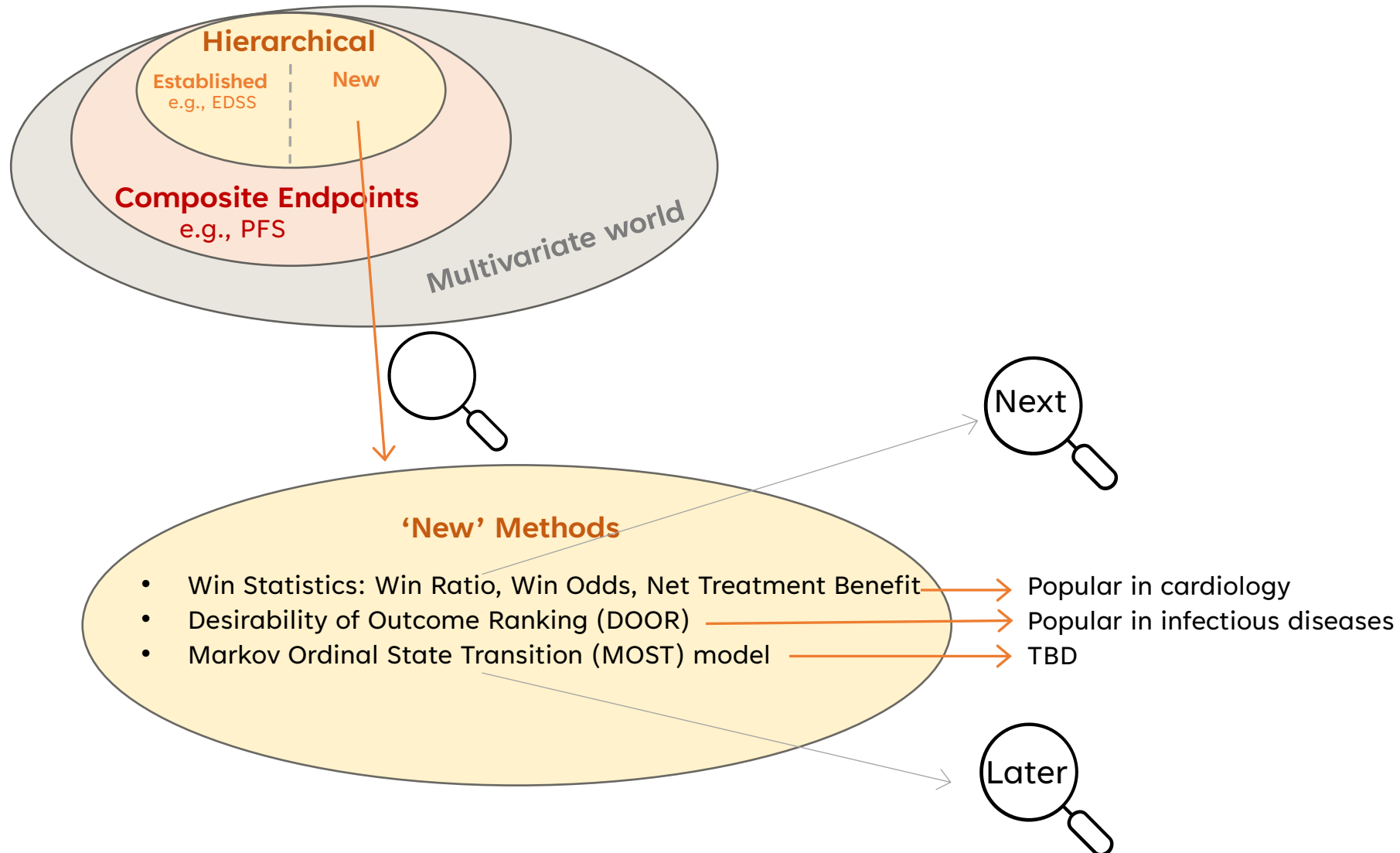
A WORLD INSIDE A WORLD: LANDSCAPE OF HCEs




A WORLD INSIDE A WORLD: LANDSCAPE OF HCEs



A WORLD INSIDE A WORLD: LANDSCAPE OF HCEs





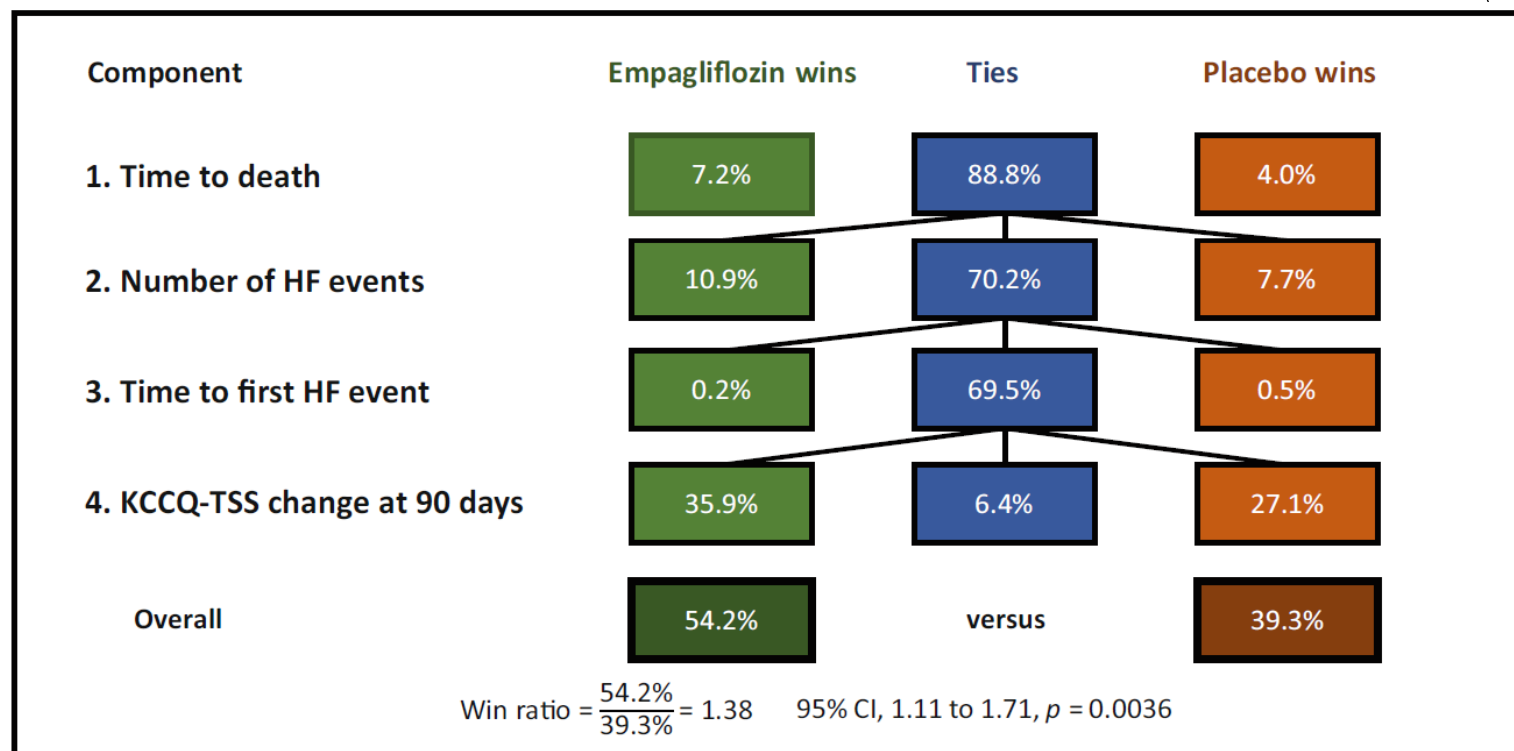
3. Win Statistics: the Good, the Bad and the Misleading

WHAT IS A WIN STATISTIC

What it is

- Summary of treatment effect across the HCE hierarchy.
- Method:
 - Compare each treated subject to each control subject.
 - Determine a “win” based on the most severe outcome with a difference.
 - If tied, move to the next component in the hierarchy.
- Commonly used: Win Ratio (WR), Win Odds (WO), Net Benefit (NB)

Illustration



Building blocks: $P(Y_i^A > Y_j^C)$ and $P(Y_j^C > Y_i^A)$

$$\text{Win Ratio} = P(Y_i^A > Y_j^C) / P(Y_j^C > Y_i^A)$$

$$\text{Net Benefit} = P(Y_i^A > Y_j^C) - P(Y_j^C > Y_i^A)$$

ESTIMAND (ICH E9 ADDENDUM)

Treatment: Defines the treatment and alternative treatment of interest, including choices of standard-of-care.

Population: Identifies the group of subjects relevant to the clinical question.

Variable (Endpoint): specifies the measurement or outcome used to address the clinical question.

Intercurrent Event Strategy:
Describes how intercurrent events are accounted for, e.g. through treatment policy, hypothetical, or composite strategies.

Population-level Summary:
Details the summary measure of the variable across the population

- Demonstrating the existence of treatment effects **and quantifying their magnitude**.
- Causal comparison of the outcome with the intervention to the outcome that would have occurred **for the same subjects** under an alternative intervention.
- If a treatment on average leads to a higher win probability compared to a comparator, this indicates the existence of a positive treatment effect. However, it is more complex to claim that the estimated WR also answers the "how much better" question.

CAUSAL INTERPRETATION

HANDS PARADOX

Toy example

Factual and counterfactual responses for three subjects under two treatments.

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

CAUSAL INTERPRETATION

HANDS PARADOX

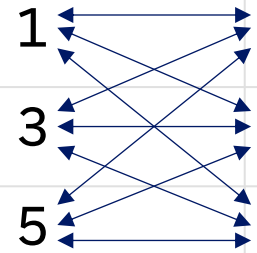
Toy example

Factual and counterfactual responses for three subjects under two treatments.

WR (**population-level**), comparing all outcomes in the two arms is:

$$\frac{(0+0+0) + (0+1+0) + (0+1+1)}{(1+1+1) + (1+0+1) + (1+0+0)} = 0.5$$

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



CAUSAL INTERPRETATION

HANDS PARADOX

Toy example

Factual and counterfactual responses for three subjects under two treatments.

WR (**population-level**), comparing all outcomes in the two arms is:




$$\frac{(0+0+0) + (0+1+0) + (0+1+1)}{(1+1+1) + (1+0+1) + (1+0+0)} = 0.5$$

WR (**individual-level**)

$$\frac{(0+1+1)}{(1+0+0)} = 2$$

Thus, not only are they different, but they are pointing in opposite directions

This individual-level is not identifiable in randomized experiment

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

NON-TRANSITIVITY

Toy example (Efron's dice)

Three treatments, A, B, and C.

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

So, A better than B, better than C, better than A.

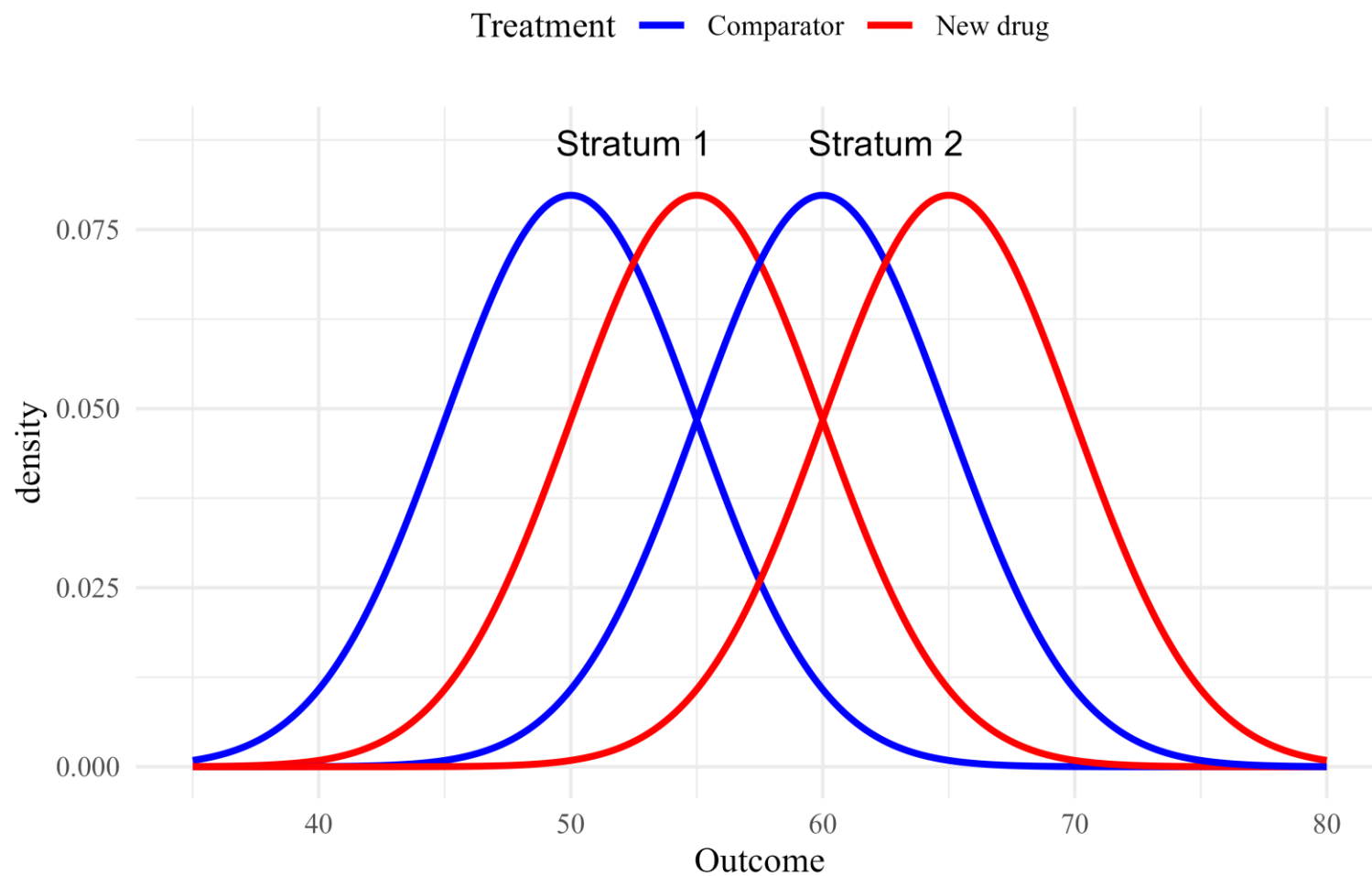
A	B	C
2	1	3
4	6	5
9	8	7

NON-COLLAPSIBILITY

Stratum	% of Population	New Drug (Mean)	Comparator (Mean)	WR
Strata 1	50%	55	50	3.17
Strata 2	50%	65	60	3.17
Combined	100%	–	–	2.18

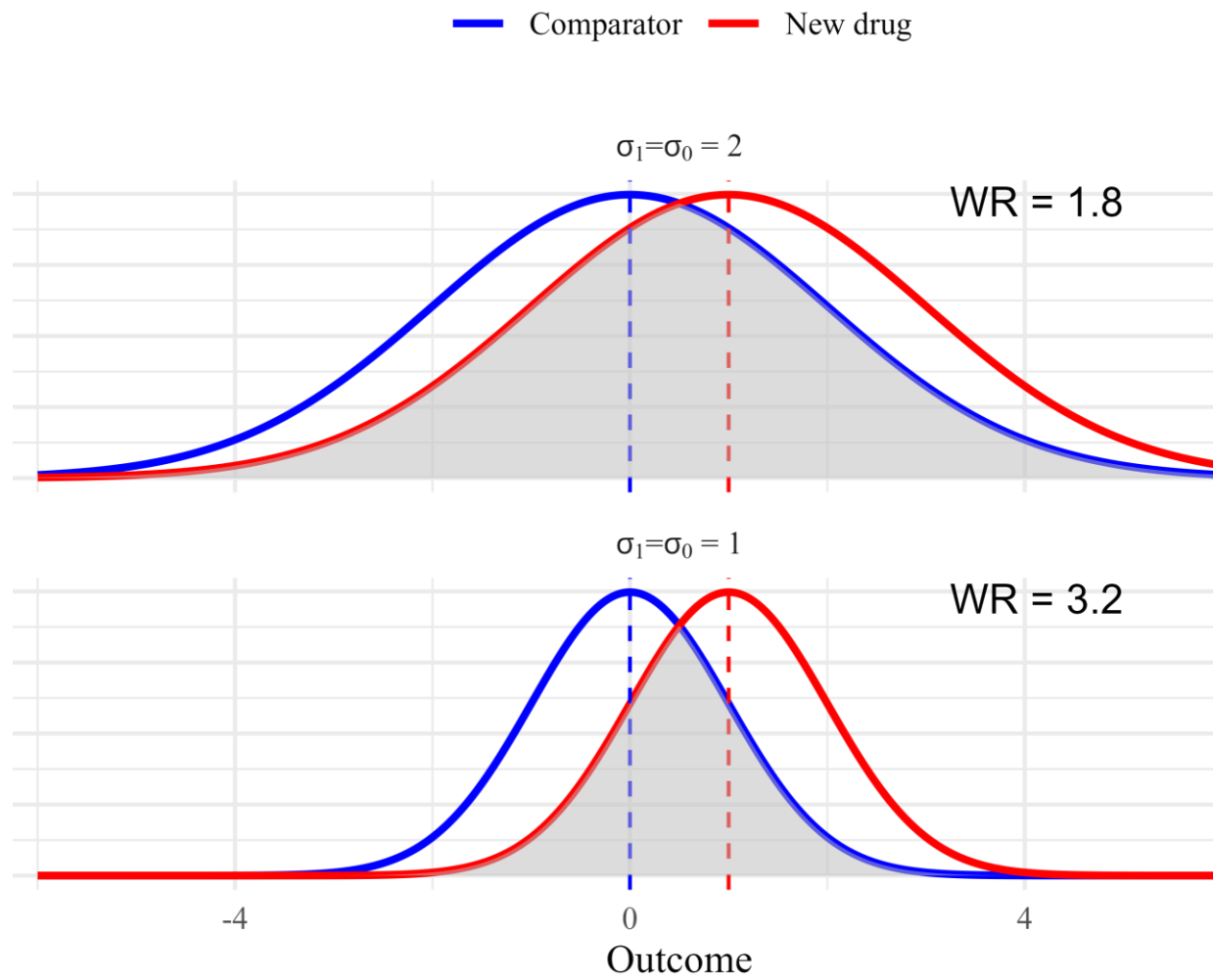
normally distributed response in each stratum, a common standard deviation of 5

NON-COLLAPSIBILITY



normally distributed response in each stratum, a common standard deviation of 5

VARIANCE DEPENDENCE



DISCUSSION

Estimand

- Defining an appropriate estimand is challenging and crucial for accurately reflecting the clinical question.

Dependence on Variance

- The WR's value is influenced by the variance of continuous components in HCEs, making it predominantly a measure of discriminatory character between active and control groups rather than an effect measure.

Causal Interpretation Challenges

- The WR's non-collapsible nature complicates causal interpretations and makes comparisons across different trials or in meta-analyses problematic.
- The Hand's paradox illustrates that the WR can exhibit contrasting effects at the population-level versus the individual-level.
- Non-transitivity potentially complicates treatment comparisons.

Recommendations

- While the WR might be useful for establishing treatment effects, its interpretation requires awareness of its limitations.
- Suggests treating WR as a discriminatory measure, like non-parametric tests, and acknowledging its challenges, particularly in defining relevant estimands and causal interpretations.



4. The Baby and the Bathwater

THE BABY AND THE BATHWATER



Me: the good cop

Two possible routes

- i. Build on Win Statistics : help improve interpretation, communication, technical aspects
- ii. Build on the spirit of Win Statistics: explore (one of the) alternatives in the realm of HCEs

THE BABY AND THE BATHWATER



Me: the good cop

Two possible routes

- i. Build on Win Statistics : help improve interpretation, communication, technical aspects
- ii. Build on the spirit of Win Statistics: explore (one of the) alternatives in the realm of HCEs

Observation

- *The good*: Win Statistics attempt to
 - Account for multidimensional aspect of patient's experience,
 - While acknowledging order of desirability

THE BABY AND THE BATHWATER



Me: the good cop

Two possible routes

- i. Build on Win Statistics : help improve interpretation, communication, technical aspects
- ii. Build on the spirit of Win Statistics: explore (one of the) alternatives in the realm of HCEs

Observation

- *The good*: Win Statistics attempt to
 - Account for multidimensional aspect of patient's experience,
 - While acknowledging order of desirability
- *The complex*: Win Statistics are built on $P(Y_i^A > Y_j^C)$
 - Technical & interpretational challenges,
 - *Inevitably* affected by time: in long trials death will 'dominate', not in short trials

THE BABY AND THE BATHWATER



Me: the good cop

Two possible routes

- i. Build on Win Statistics : help improve interpretation, communication, technical aspects
- ii. Build on the spirit of Win Statistics: explore (one of the) alternatives in the realm of HCEs

Observation

- *The good*: Win Statistics attempt to
 - Account for multidimensional aspect of patient's experience,
 - While acknowledging order of desirability
- *The complex*: Win Statistics are built on $P(Y_i^A > Y_j^C)$
 - Technical & interpretational challenges,
 - *Inevitably* affected by time: in long trials death will 'dominate', not in short trials

Ambition

Enrich* the world of HCEs:

- Give opportunities to go beyond $P(Y_i^A > Y_j^C)$: HCEs are larger than these probabilities
- In doing so, refine the discussion about the 'when'

* Based on other people's work: F. Harrell, M. Shun-Shin, and a lot of very smart colleagues

BACK TO FIRST EXAMPLE

Potential alternatives

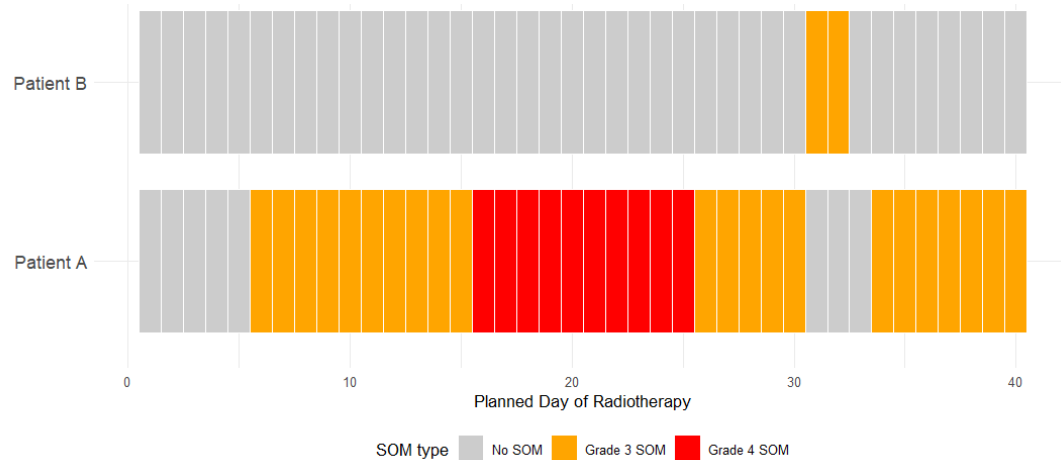
- Multivariate considerations are inevitably affected by time
- Put them in plain sight, e.g.,
 - Endpoints expressed as (composite) expected times
 - Endpoints expressed as (composite) milestone probabilities

BACK TO FIRST EXAMPLE

Potential alternatives

- Multivariate considerations are inevitably affected by time
- Put them in plain sight, e.g.,
 - Endpoints expressed as (composite) expected times
 - Endpoints expressed as (composite) milestone probabilities

Back to the first example

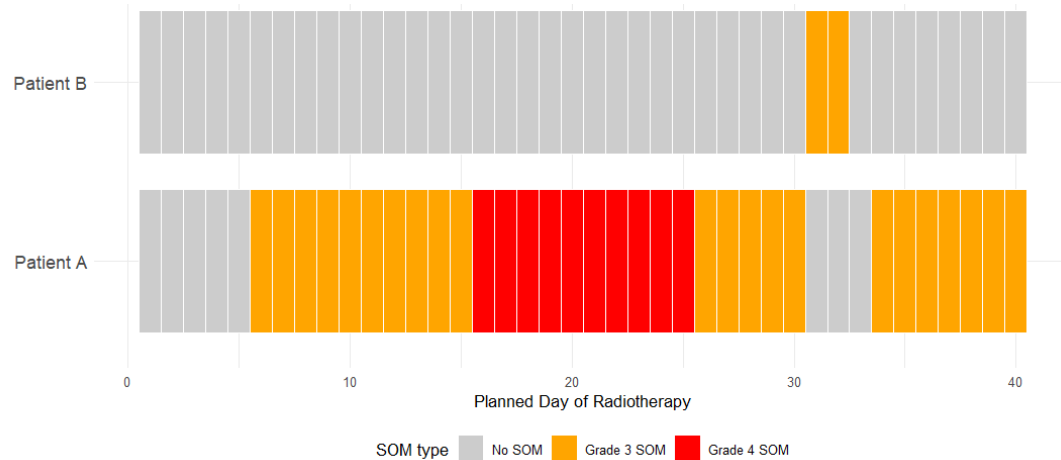


BACK TO FIRST EXAMPLE

Potential alternatives

- Multivariate considerations are inevitably affected by time
- Put them in plain sight, e.g.,
 - Endpoints expressed as (composite) expected times
 - Endpoints expressed as (composite) milestone probabilities

Back to the first example



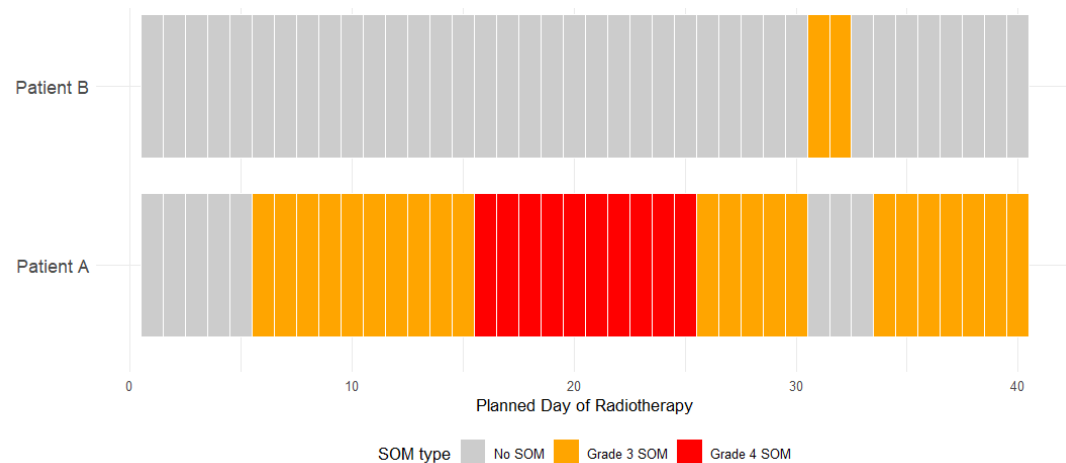
Spirit

- i. Model *raw* data, respect timing and severity of events: discrete-time multistate process
- ii. Extract estimator for the estimand of interest: model i. is a means to an end

BACK TO FIRST EXAMPLE

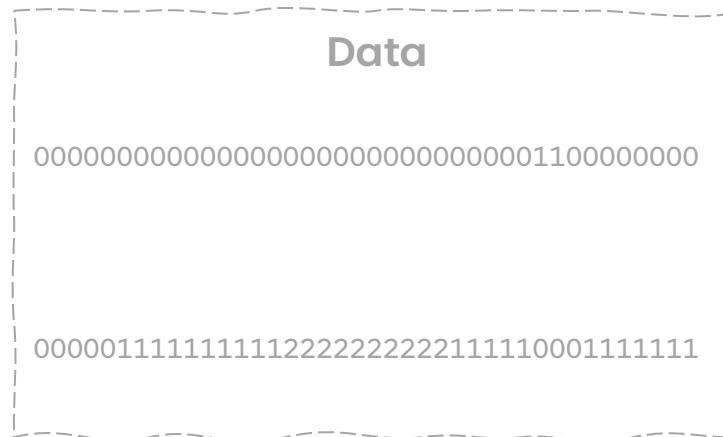
Process

1. The (oversimplified) 'journeys'



Process

2. The numerical translation



46

Process

The chart displays the planned day of radiotherapy for two patients, Patient A and Patient B, categorized by SOM type. The x-axis represents the 'Planned Day of Radiotherapy' from 0 to 40. The y-axis lists the patients. The legend indicates three SOM types: No SOM (gray), Grade 3 SOM (yellow), and Grade 4 SOM (red).

Patient	SOM type	Planned Day of Radiotherapy (Approximate)
Patient B	No SOM	0 - 31
	Grade 3 SOM	31 - 33
Patient A	No SOM	0 - 5
	Grade 3 SOM	5 - 15
	Grade 4 SOM	15 - 25
	Grade 3 SOM	25 - 31

Data

```
0000011111111122222222211111000111111
```

Grading events on each unit of (short) time – easier consensus

Model ordinal longitudinal data, e.g.,
using a first-order discrete-time Markov
(partial) proportional odds model

Process

The chart displays the planned day of radiotherapy for two patients, Patient A and Patient B, categorized by the type of Secondary Osteomyeloma (SOM). The x-axis represents the 'Planned Day of Radiotherapy' from 0 to 40. The y-axis lists the patients. The legend indicates three SOM types: No SOM (grey), Grade 3 SOM (yellow), and Grade 4 SOM (red).

Patient	SOM type	Planned Day of Radiotherapy (Approximate)
Patient B	No SOM	0 - 31
	Grade 3 SOM	31 - 33
Patient A	No SOM	0 - 5
	Grade 3 SOM	5 - 15
	Grade 4 SOM	15 - 25
	Grade 3 SOM	25 - 31

[illegible]

Grading events on
each unit of (short)
time – easier
consensus

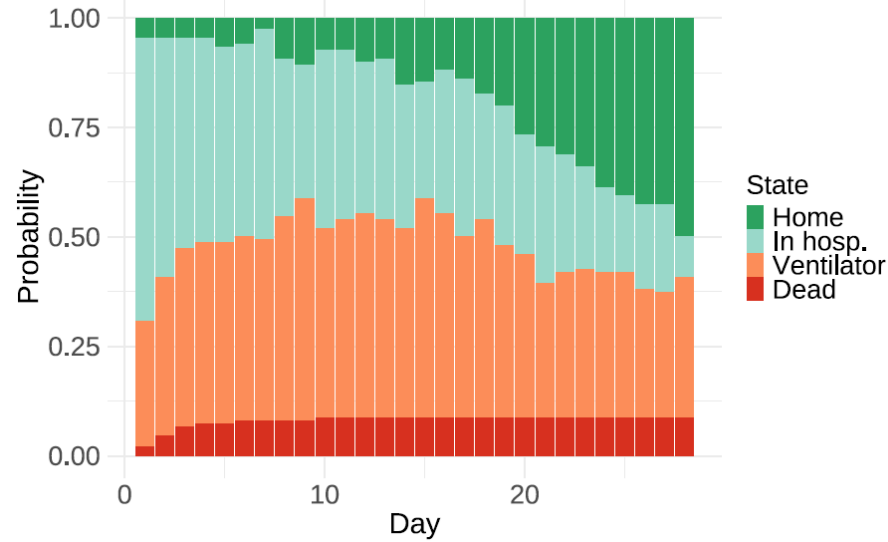
- Expected times:
 - Mean time ‘unwell’ (SOM Gr ≥ 3) throughout the trial
- Landmark probabilities:
 - P(SOM-free at day x without having spent an excessive amount of days with SOM Gr ≥ 3)

Model ordinal longitudinal data, e.g.,
using a first-order discrete-time Markov
(partial) proportional odds model

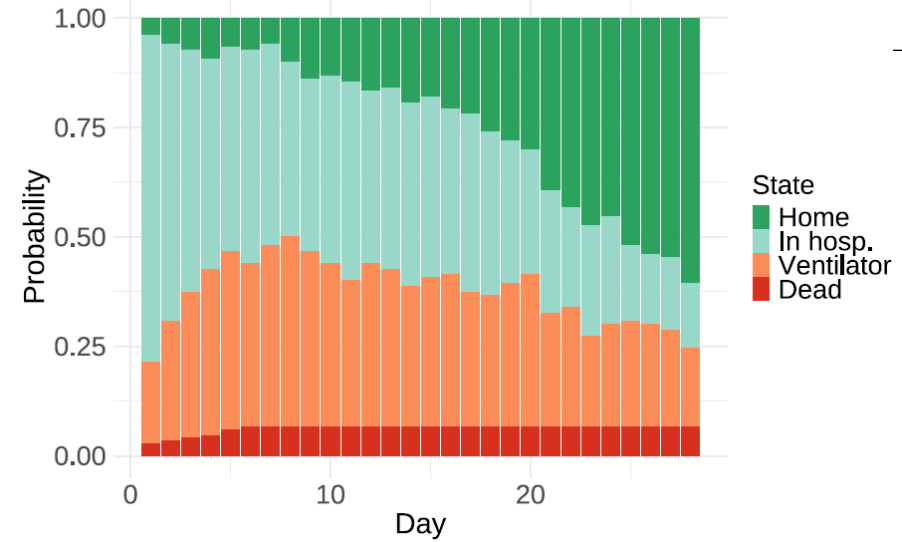
ANOTHER EXAMPLE

COVID-19 Trial

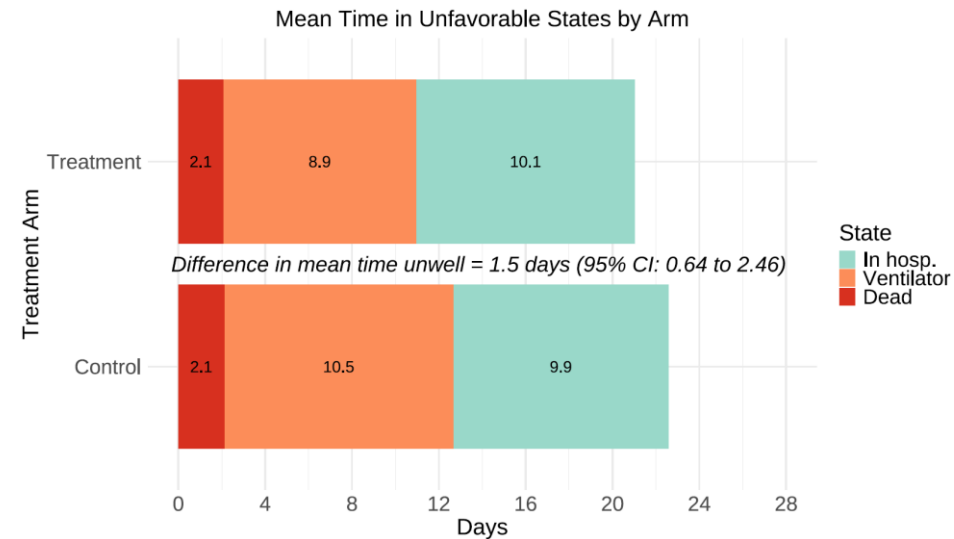
Daily State Occupancy Probabilities - Control



Daily State Occupancy Probabilities - Treatment

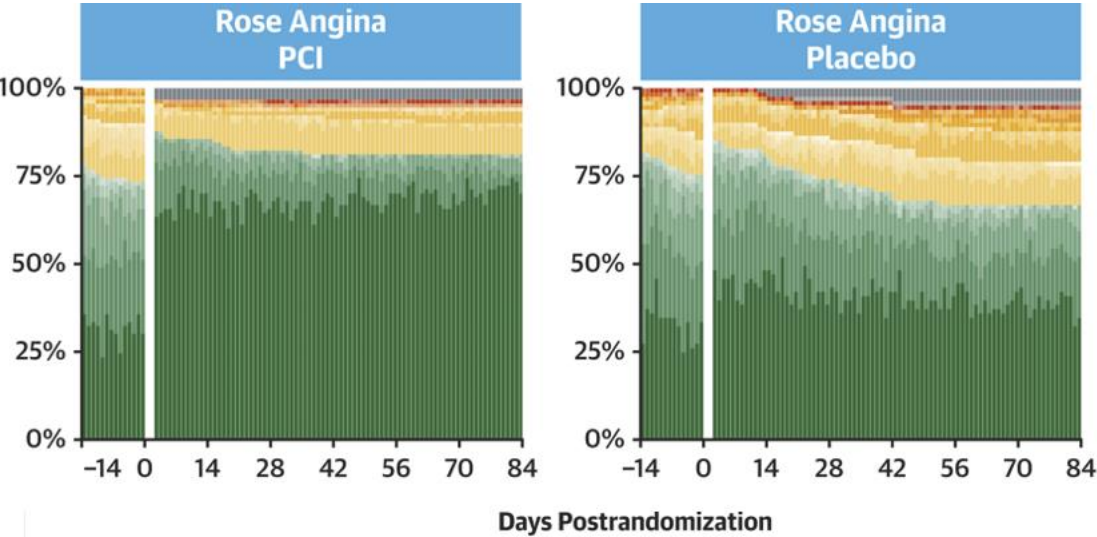
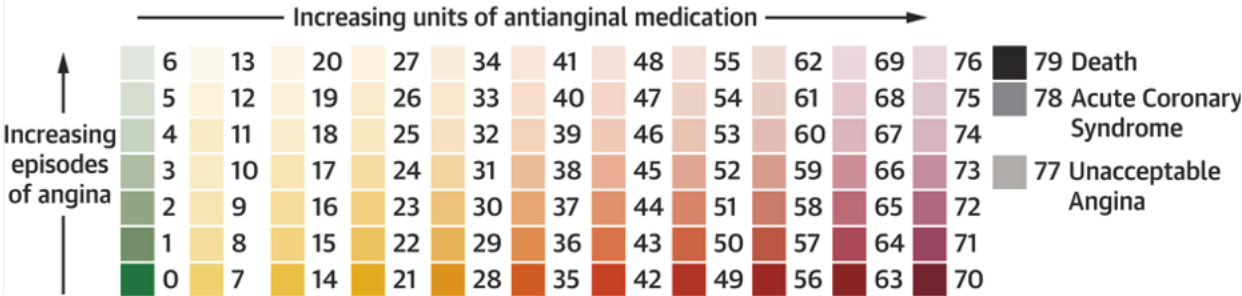
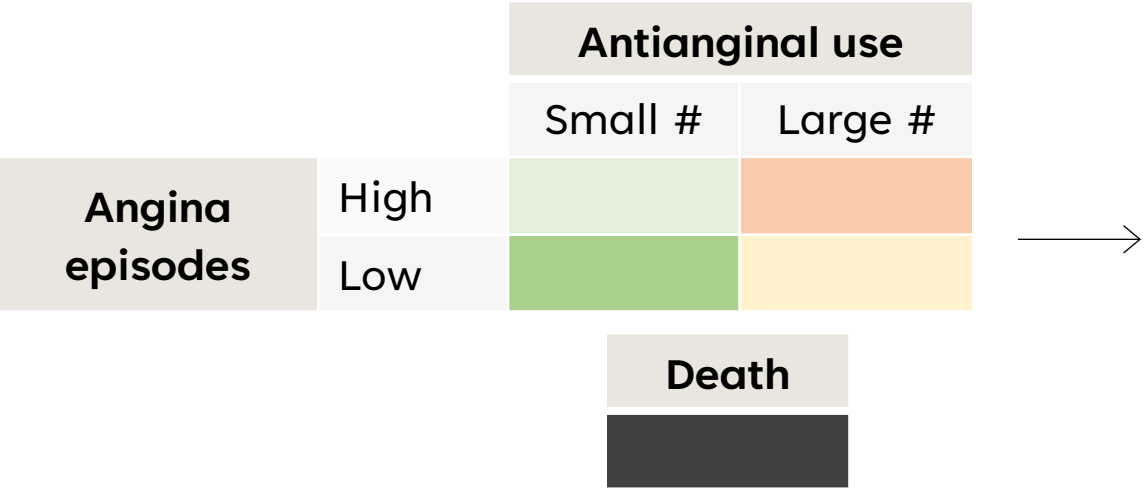



One possible contrast



BACK TO THE SECOND EXAMPLE

From simple to less simple





5. Built-In Tensions and a Personal Hope

BUILT-IN TENSIONS

Multivariate by ambition, complex by accident

- Summary can correspond to very different underlying realities
- Can obtain 'overall' claim without being able to pinpoint the origin
- Temporal maturity of components may vary
- Subjectivity in ordering?

BUILT-IN TENSIONS

Multivariate by ambition, complex by accident

- Summary can correspond to very different underlying realities
- Can obtain 'overall' claim without being able to pinpoint the origin
- Temporal maturity of components may vary
- Subjectivity in ordering?

Is this any different from 'traditional' composite endpoints?

BUILT-IN TENSIONS

Multivariate by ambition, complex by accident

- Summary can correspond to very different underlying realities
- Can obtain 'overall' claim without being able to pinpoint the origin
- Temporal maturity of components may vary
- Subjectivity in ordering?

Is this any different from 'traditional' composite endpoints?

A PERSONAL HOPE

Building towards concrete criteria for agreeable combination

A suggestion for the 'when':



BUILT-IN TENSIONS

Multivariate by ambition, complex by accident

- Summary can correspond to very different underlying realities
- Can obtain 'overall' claim without being able to pinpoint the origin
- Temporal maturity of components may vary
- Subjectivity in ordering?

Is this any different from 'traditional' composite endpoints?

A PERSONAL HOPE

Building towards concrete criteria for agreeable combination

A suggestion for the 'when':

1. Gains from multivariate outweigh complexity
 - a. Composite strategy for ICEs
 - b. Benefit-risk in a single analysis
2. Signal is objectively weak***



*** No p-values were harmed in the making of these stars – just a dramatic warning for nuance

A PERSONAL HOPE

A suggestion for the 'when':

1. Gains from multivariate outweigh the added complexity
 - a. Composite strategy for ICEs
 - b. Benefit-risk in a single analysis
2. Signal is objectively weak

Components should satisfy	Composite	HCEs
Shared biological mechanism	✓	✓
Similar level of objectivity in measurement	✓	✓
No dominance by 'lesser' component	✓	✓
Consistent direction of effect	✓	✓◇
Similar clinical relevance	✓	

Principles of composite endpoints should not be forgotten

By construction 'no', but... where do we place the cursor?



WHAT THIS TALK WAS ABOUT

Under the umbrella: hierarchical composite endpoints form a landscape

This is new, this is not new

Multivariate: Everything Everywhere All At Once

WHAT THIS TALK WAS ABOUT

Under the umbrella: hierarchical composite endpoints form a landscape

This is new, this is not new

Multivariate: Everything Everywhere All At Once

Is this a case where “perfect is the enemy of the good”?



THANK YOU

Henrik F. Thomsen & Mickaël De Backer

