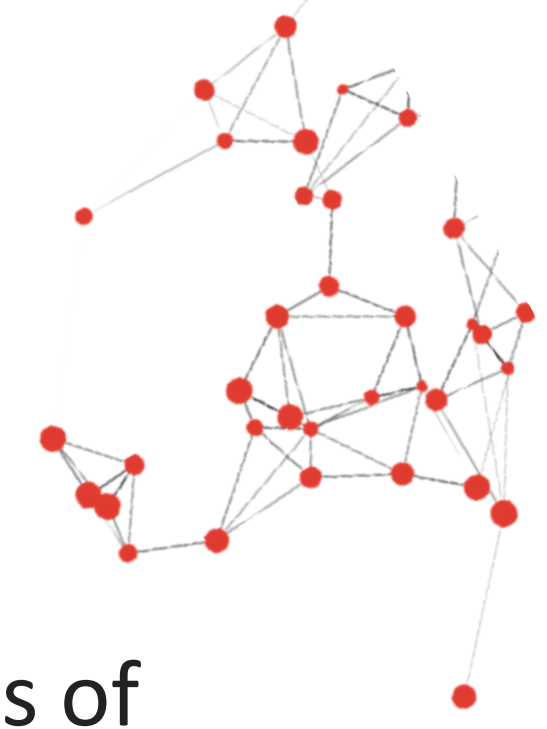


Generalized pairwise comparisons: A practical guide to the design and analysis of patient-centric trials

Johan Verbeeck PhD

johan.verbeeck@uhasselt.be

Data Science Institute

UHasselt - Belgium

WWW.UHASSELT.BE/DSI

UNIVERSITY OF COPENHAGEN

Brice Ozenne PhD

brice.ozenne@nru.dk

Biostatistics & Neurobiology Research Unit

University of Copenhagen- Denmark

Author disclosures

- We declare no conflicts of interest

Agenda

- Motivating examples
  1. Time-to-first event
  2. Inherently multivariate outcomes
  3. Multivariate outcomes of different types
  4. Benefit-risk assessment
- Concept of GPC

1:00-2:00

Agenda

- Motivating examples
  1. Time-to-first event
  2. Inherently multivariate outcomes
  3. Multivariate outcomes of different types
  4. Benefit-risk assessment
- Concept of GPC

1:00-2:00

- Inference for GPC and software
- Examples, revisited

2:30-3:30

Agenda

- Motivating examples
  1. Time-to-first event
  2. Inherently multivariate outcomes
  3. Multivariate outcomes of different types
  4. Benefit-risk assessment

1:00-2:00

- Concept of GPC
- Inference for GPC and software
- Examples, revisited

2:30-3:30

- Advanced Topics
  1. Censoring
  2. Stratification
  3. Covariate adjustment
- Trial design

4:00-5:00

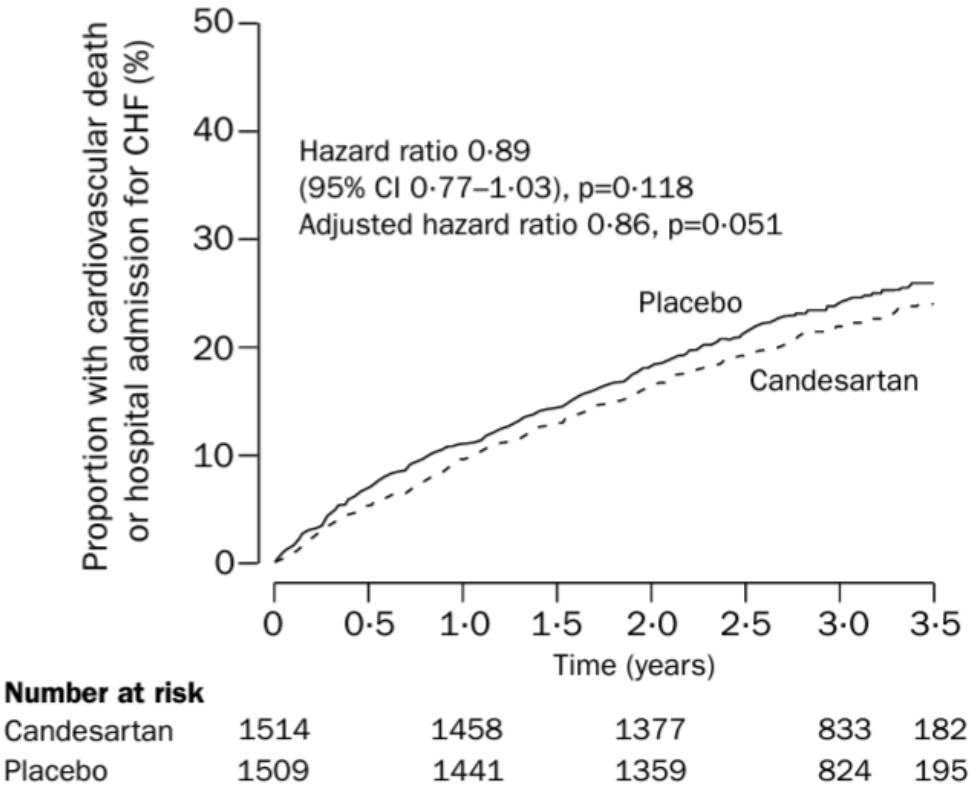
Motivating examples

1. Composite of survival outcomes

Effects of candesartan in patients with chronic heart failure and preserved left-ventricular ejection fraction: the CHARM-Preserved Trial

Salim Yusuf, Marc A Pfeffer, Karl Swedberg, Christopher B Granger, Peter Held, John J V McMurray, Eric L Mic
*Bertil Olofsson, Jan Östergren, for the CHARM Investigators and Committees**

	Candesartan (n=1514)	Placebo (n=1509)
Cardiovascular death or hospital admission for CHF	333 (22·0%)	366 (24·3%)
Cardiovascular death	170 (11·2%)	170 (11·3%)
Hospital admission for CHF	241 (15·9%)	276 (18·3%)



Yusuf et al. Lancet (2003)

Issues with time-to-first event analyses

	Candesartan (n=1514)	Placebo (n=1509)
Cardiovascular death or hospital admission for CHF	333 (22·0%)	366 (24·3%)
Cardiovascular death	170 (11·2%)	170 (11·3%)
Hospital admission for CHF	241 (15·9%)	276 (18·3%)

→

Events in time-to-first composite

Candesartan	Placebo
92 (54%)	90 (53%)
241 (100%)	276 (100%)

46% of CV deaths are ignored

Yusuf et al. Lancet (2003)
Pocock et al. EHJ (2012)

Issues with time-to-first event analyses

- Emphasis is on time of event, rather than severity of event
- Ignores repeated events

	Candesartan (n=1514)	Placebo (n=1509)
Number of patients (%)		
None	1284 (84·8%)	1230 (81·5%)
1	132 (8·7%)	157 (10·4%)
2	54 (3·6%)	59 (3·9%)
⩾3	44 (2·9%)	63 (4·2%)
Number of patients admitted to hospital (number of admissions)	230 (402)	279 (566)

*Investigator reported, with CHF as primary reason (p=0·014 for distribution).

Table 3: **Numbers of hospital admissions for worsening heart failure***

Verbeeck et al. JACC (2023)
Verbeeck et al. EHJ:ACVC (2024)
Yusuf et al. Lancet (2003)

Issues with hazard ratio

- Misinterpretation:

Example			
	Event	No event	
Control	A	B	A total of n_1 patients followed for a cumulative time t_1
Active	C	D	A total of n_2 patients followed for a cumulative time t_2

Risk interpretation		
	Interpretation	Formula
Hazard ratio	Instantaneous risk reduction or Relative rate reduction	$\frac{A/t_1}{C/t_2}$
Risk ratio	Relative risk reduction	$\frac{A/n_1}{C/n_2}$
Risk difference	Absolute risk difference	$A/n_1 - C/n_2$

- HR is time-dependent unless the hazard rates are proportional over time

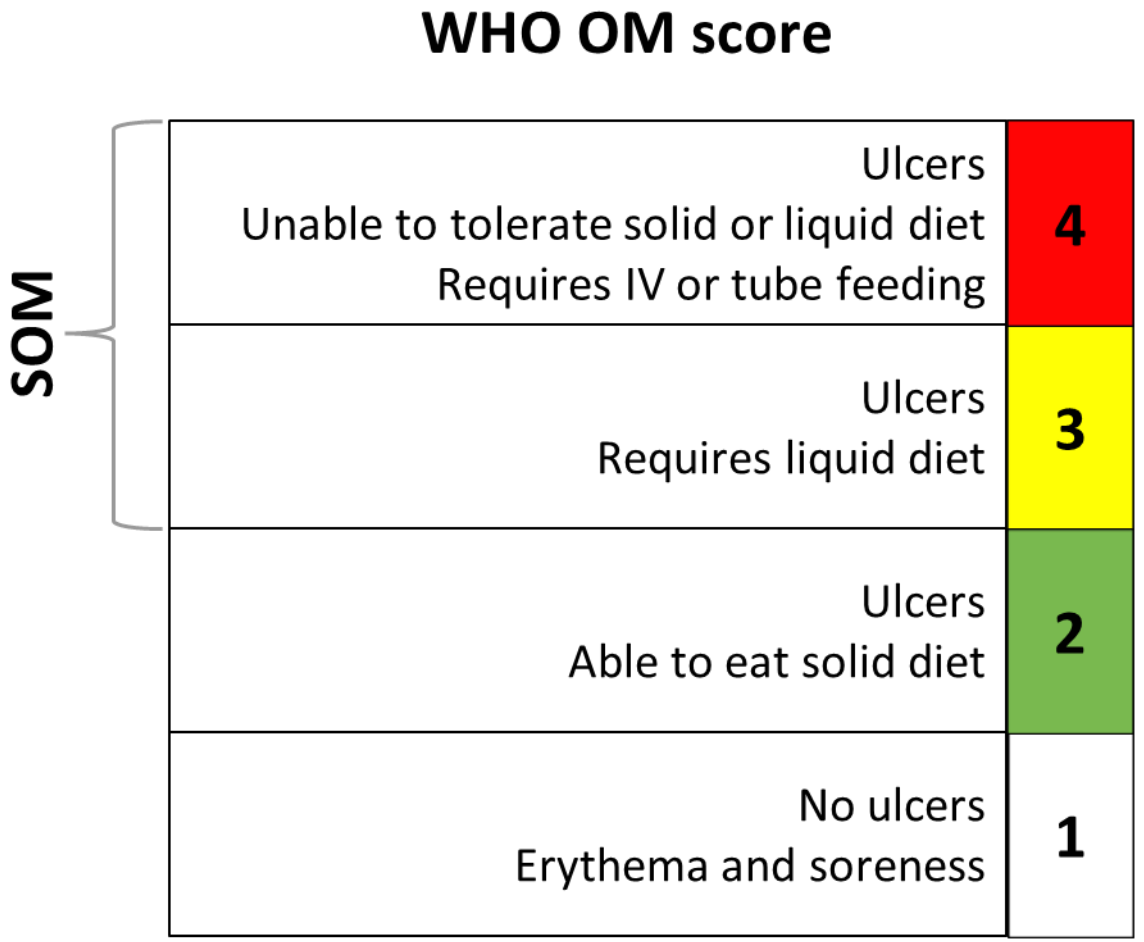
Verbeeck et al. JACC (2023)
Verbeeck et al. EHJ:ACVC (2024)

2. Inherently multivariate outcomes

- Patients with locally advanced head & neck cancers may be treated with radiotherapy and cisplatin
- Two thirds of patients develop severe oral mucositis (SOM)
- SOM is a highly patient-relevant toxicity that
  - impacts quality of life
  - may result in treatment delays
- There is currently no FDA-approved drug for the prevention of SOM

Anderson et al. (Submitted, 2024)

Severe Oral Mucositis



Traditional analysis

- Compare incidence of SOM in patients receiving treatment vs. placebo
- Example : ROMAN trial compared 241 patients treated with avasopasem manganese (AVA) vs. 166 patients receiving placebo (PBO)
  - Incidence of SOM : 64% PBO vs. 54% AVA
  - Absolute reduction in risk = 10%
  - *P*-value = 0.045
- But…

Anderson et al. (Submitted, 2024)

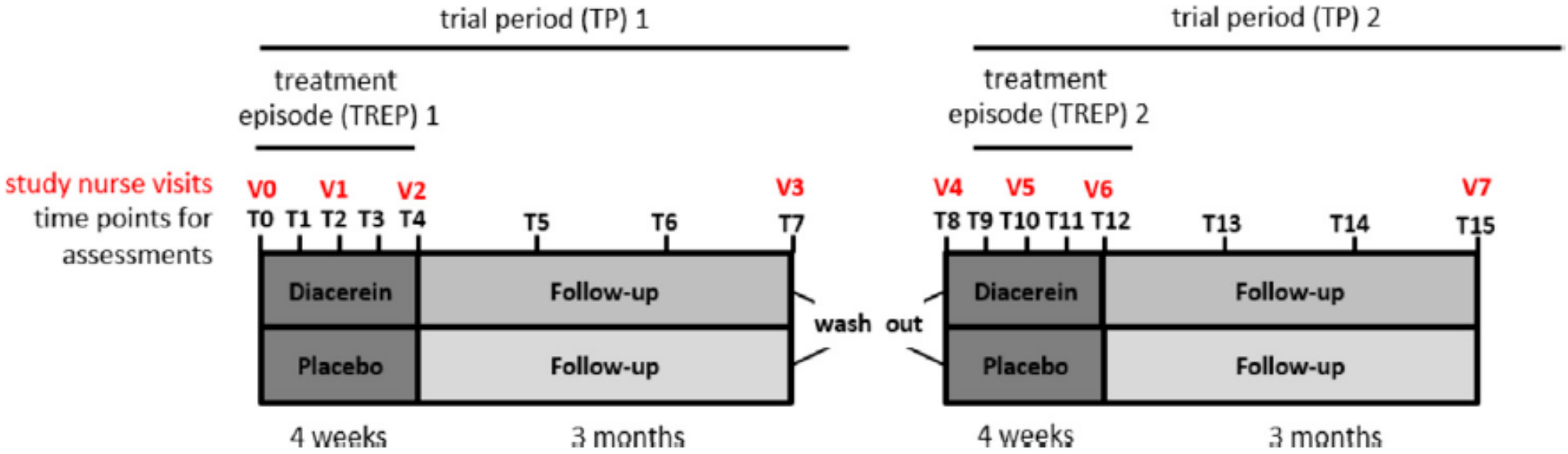
Traditional analysis

- Comparison of SOM incidence ignores
  - SOM severity (grade 4 much worse than grade 3)
  - SOM duration (from a couple of days to more than a month)
  - SOM onset time (later preferable because lower impact on radiotherapy)
- SOM is *inherently* a multivariate outcome
- Ignoring key features of SOM results in
  - loss of power
  - loss of clinical interpretability

Anderson et al. (Submitted, 2024)

3. Multivariate outcomes of different types

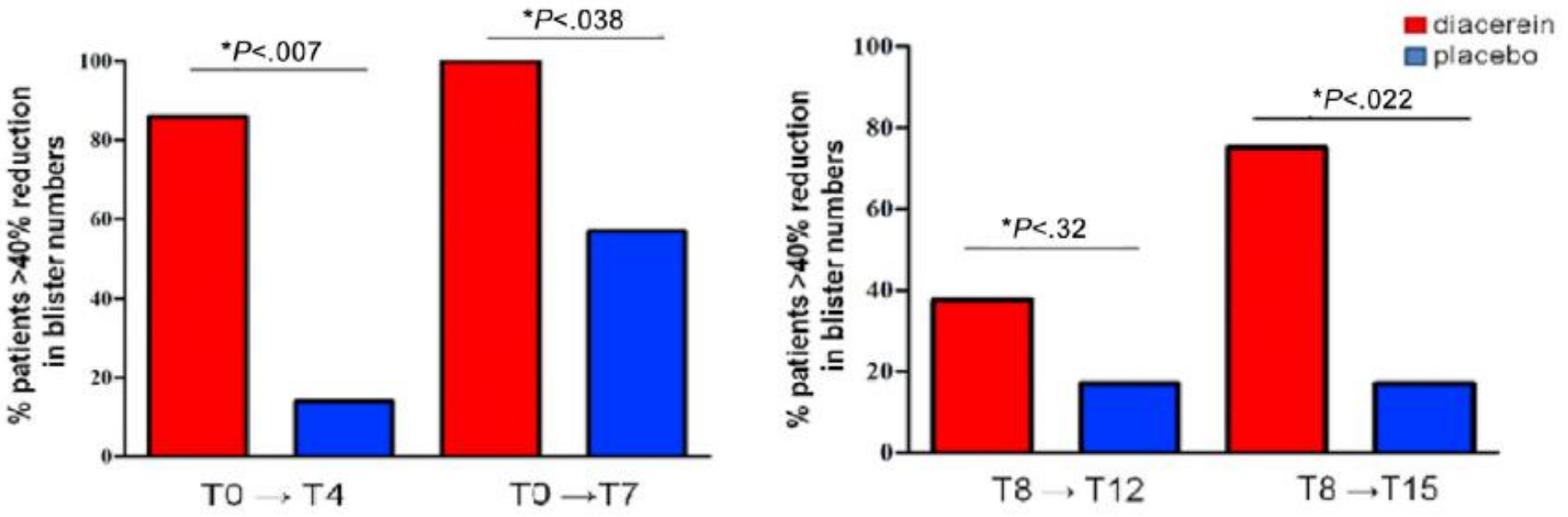
- Rare skin disease: Epidermolysis bullosa simplex
- Formation of blisters under low mechanical stress
- 16 pediatric subjects treated with placebo and diacerin cream in a longitudinal cross-over trial



Wally et al. . J Am Acad Dermato (2018)

Inconclusive results - primary endpoint analysis

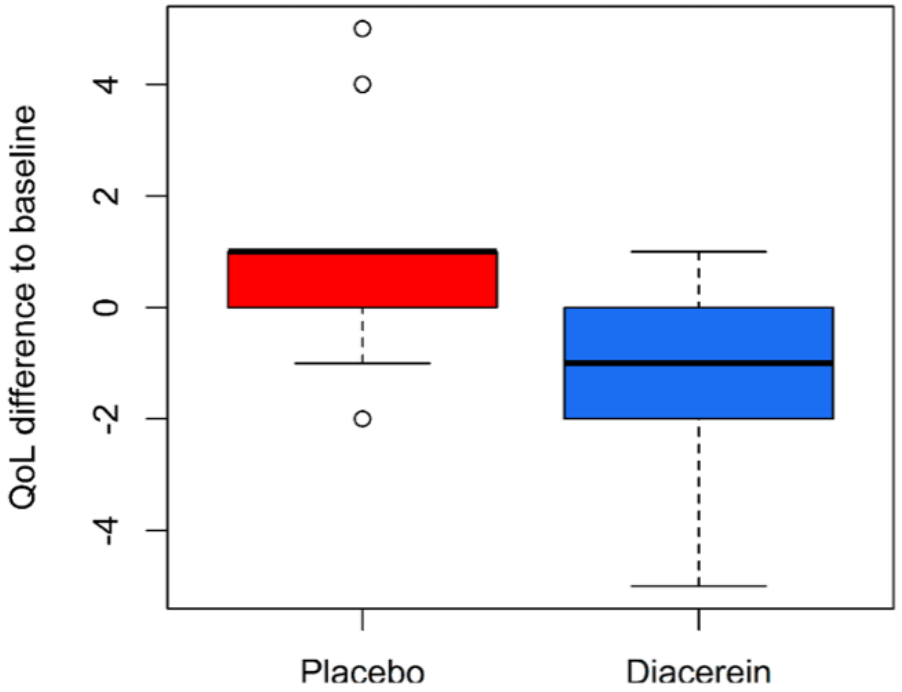
- Primary endpoint:
  - >40% reduction in blisters compared to baseline (binary outcome) at week 4
  - Barnard test (~Fisher exact test 2x2 table) per treatment period

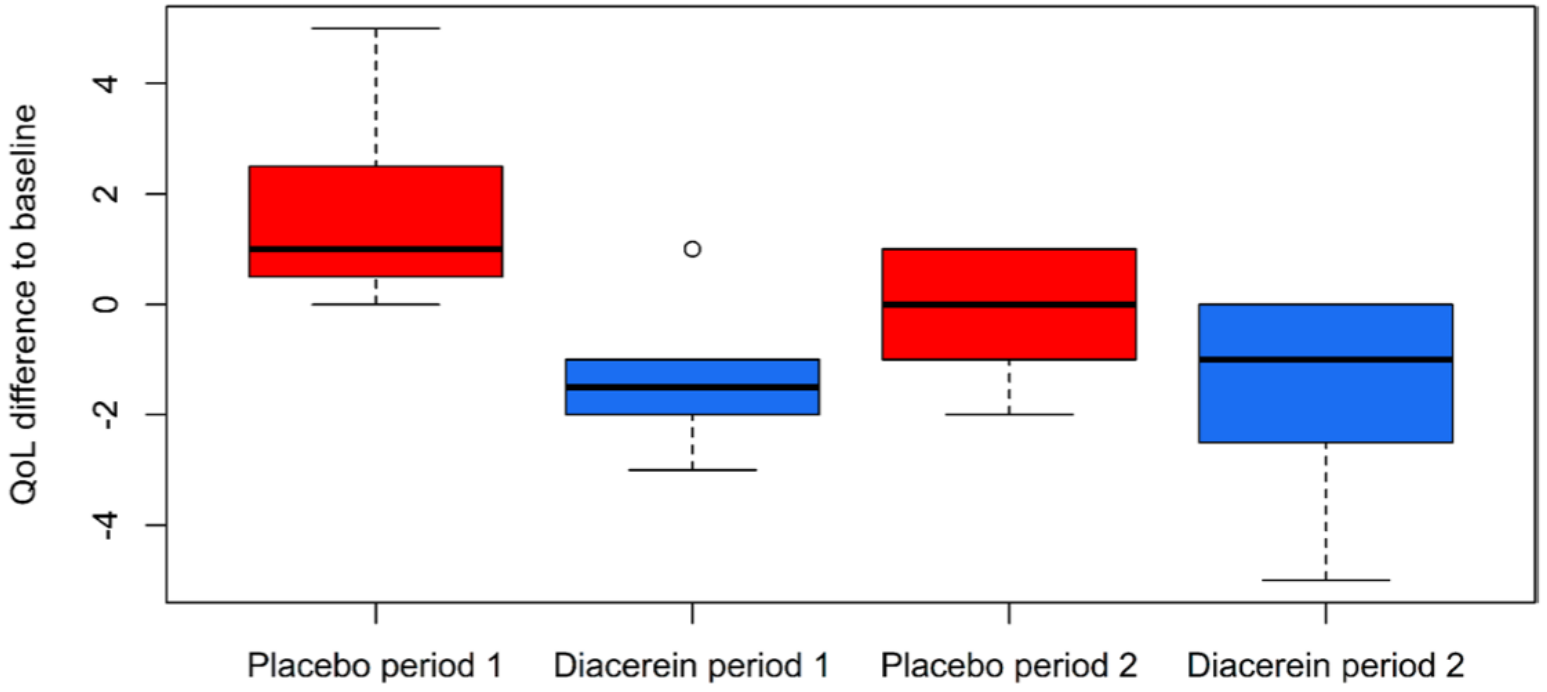


Wally et al. . J Am Acad Dermato (2018)

Patient-centric outcome ignored

- Formation of blisters under low mechanical stress – affects QoL





How to combine information from QoL with blister information?

Wally et al. . J Am Acad Dermato (2018)
Verbeeck et al. *OJRD* (2023)

4. Benefit-Risk assessment

- Simple situation :
  - binary efficacy outcome (1 = response, 0 = no response)
  - binary safety outcome (1 = no toxicity, 0 = toxicity)

Outcomes	**Treatment**	**Control**	**Difference**
Response rate (benefit)	0.5	0.2	0.3
Toxicity rate (risk)	0.6	0	0.6
Marginal benefit-risk difference			-0.3

- Naïve analysis suggests negative benefit-risk of -0.3
- How do we interpret this statistic?

Marginal Benefit-Risk analyses

- How do we interpret this statistic?
- This statistic has no interpretation
  - it ignores association between benefit and risk
  - marginal benefits and risks cannot be combined

Sensible Benefit-Risk analyses

Outcomes	**Treatment**	**Control**	**Difference**
Response rate (benefit)	0.5	0.2	0.3
Toxicity rate (risk)	0.6	0	0.6
Marginal benefit-risk difference			-0.3

In Treatment :

Toxicity	
Absent	Present
.4	.6

Response	
Present	Absent
.5	.5

Toxicity	
Absent	Present
.4	.6

Response	
Present	Absent
.5	.5

Toxicity	
Absent	Present
.4	.6

Response	
Present	Absent
.5	.5

?

No association (OR=1)

		Response	
		Present	Absent
Toxicity	Absent	.2	.2
	Present	.3	.3

Toxicity	
Absent	Present
.4	.6

Response	
Present	Absent
.5	.5

?

?

Positive association (OR= ∞)

		Response	
		Present	Absent
Toxicity	Absent	.4	.0
	Present	.1	.5

No association (OR=1)

		Response	
		Present	Absent
Toxicity	Absent	.2	.2
	Present	.3	.3

Toxicity	
Absent	Present
.4	.6

Response	
Present	Absent
.5	.5

?

?

?

Positive association (OR= ∞)

		Response	
		Present	Absent
Toxicity	Absent	.4	.0
	Present	.1	.5

Negative association (OR=0)

		Response	
		Present	Absent
Toxicity	Absent	.1	.3
	Present	.6	0

No association (OR=1)

		Response	
		Present	Absent
Toxicity	Absent	.2	.2
	Present	.3	.3

Benefit-Risk depends on association

Benefit-risk depends on the association (odds ratio, OR) between response and toxicity

- Positive association ($OR > 1$): patients who respond tend to have toxicity (*e.g.*, skin rash for inhibitors of the EGFR pathway)
- No association ($OR = 0$): response is independent of toxicity (*e.g.*, cardiac toxicities of anthracyclins)
- Negative association ($OR < 1$): patients who respond tend not to have toxicity (*e.g.*, toxicities to irinotecan in patients with enzyme deficiencies)

Concept of GPC

Wilcoxon rank-sum test

TREATMENT GROUP (T)	CONTROL GROUP (C)
$\boldsymbol{X}$ 	 $\boldsymbol{Y}$
X_1	Y_1
X_2	Y_2
...	...
X_n	Y_m

Wilcoxon. Biometrics 1945;1:80

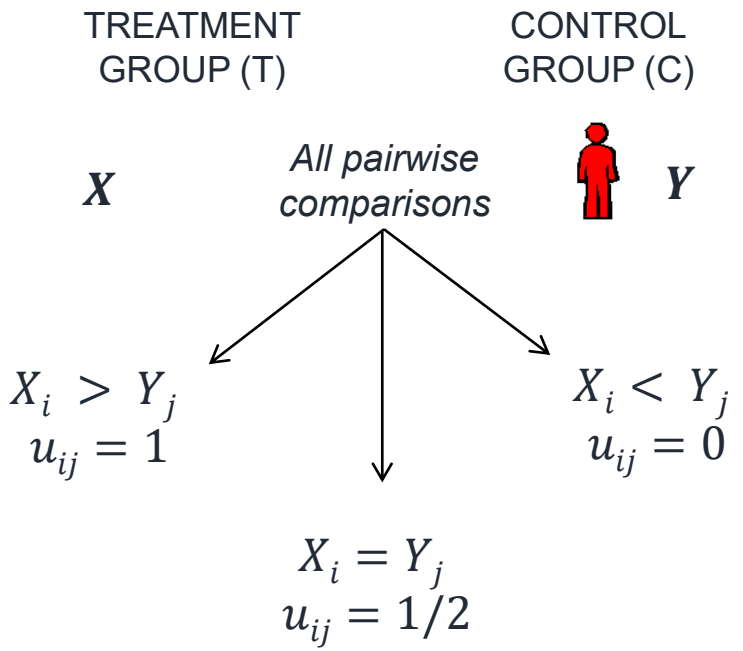
Wilcoxon rank-sum test

TREATMENT GROUP (T)	CONTROL GROUP (C)
$\boldsymbol{X}$ 	 $\boldsymbol{Y}$
X_1	Y_1
X_2	Y_2
…	…
X_n	Y_m

1. Order the $(n+m)$ elements of $\boldsymbol{X} \cup \boldsymbol{Y}$
2. Let R_i be the rank order of the i^{th} element
3. For groups of tied values, assign a rank equal to the midpoint of the unadjusted ranks
4. Calculate $U = \sum_{i=1}^{n} R_i$, the sum of ranks of the elements of $\boldsymbol{X}$
5. The statistic U has a known distribution under H_0

Wilcoxon. Biometrics 1945;1:80

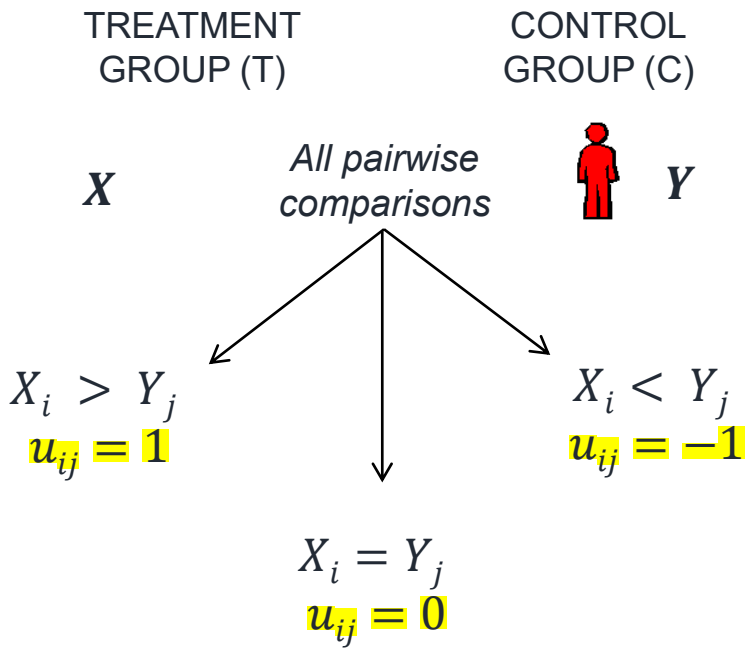
Mann-Whitney test



1. Perform pairwise comparisons between all elements of $\boldsymbol{X}$ and $\boldsymbol{Y}$
2. Calculate $u_{ij} = \begin{cases} 1\ if\ X_i > Y_j \\ 0\ if\ X_i < Y_j \\ 1/2\ \ if X_i = Y_j \end{cases}$
3. The statistic
   $U = \frac{1}{m \cdot n} \sum_{i=1}^{m} \sum_{j=1}^{n} u_{ij}$
   has a known distribution under H_0

Mann & Whitney. Ann Math Stat 1947;18:50

Mann-Whitney test

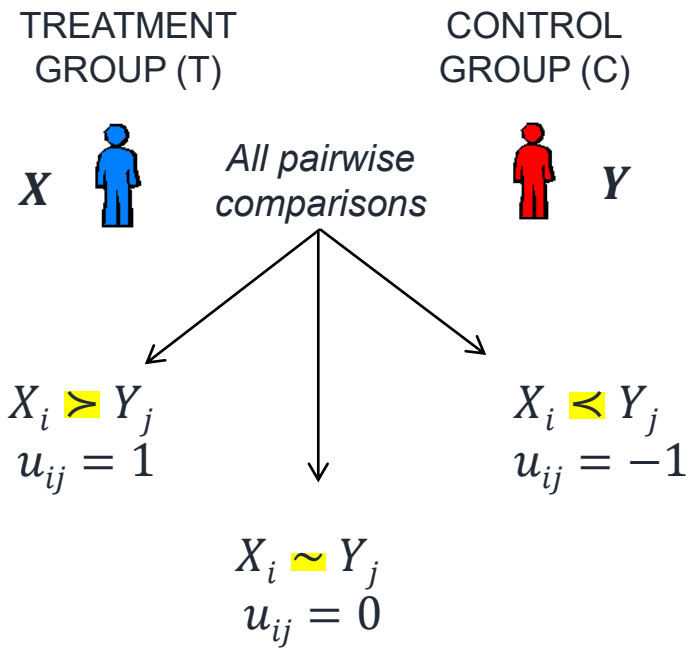


1. Perform pairwise comparisons between all elements of $\boldsymbol{X}$ and $\boldsymbol{Y}$
2. Calculate $u_{ij} = \begin{cases} 1 \; if \; X_i > Y_j \\ -1 \; if \; X_i < Y_j \\ 0 \; if X_i = Y_j \end{cases}$
3. The statistic
   $U = \frac{1}{m \cdot n} \sum_{i=1}^{m} \sum_{j=1}^{n} u_{ij}$
   has a known distribution under H_0

With these pairwise scores, U is the "Net Treatment Benefit"
$U = 0$ when there is no difference between Treatment and Control

Buyse. Stat Med 2010;29:3245

Generalized Pairwise Comparisons (GPC)



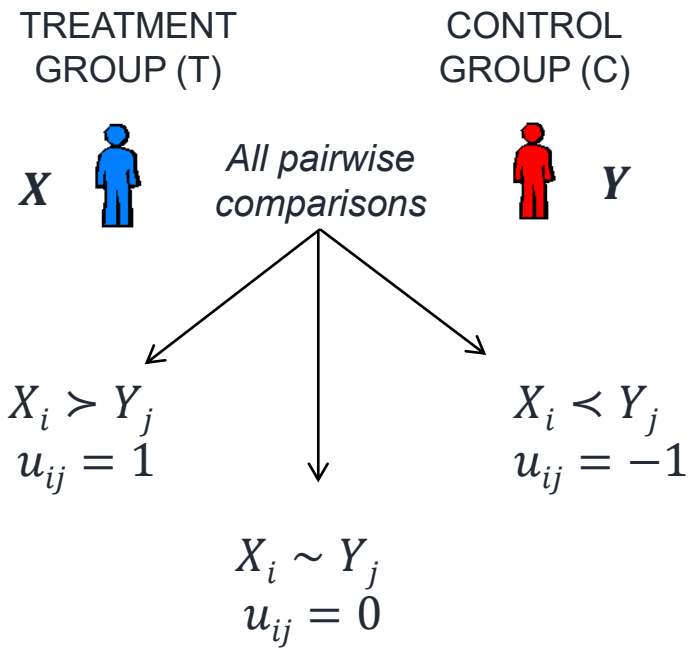
where $\succ$ stands for "better" (win)
$\prec$ stands for "worse" (loss)
$\sim$ stands for "similar" (neutral)

Note: neutral = tie or unclassifiable

1. Perform pairwise comparisons between all elements of $\boldsymbol{X}$ and $\boldsymbol{Y}$
2. Calculate $u_{ij} = \begin{cases} +1 \; if \; X_i \succ Y_j \\ -1 \; if \; X_i \prec Y_j \\ 0 \; if \; X_i \sim Y_j \end{cases}$
3. The statistic
   $U = \frac{1}{m \cdot n} \sum_{i=1}^{m} \sum_{j=1}^{n} u_{ij}$
   has a known distribution under H_0

Buyse. Stat Med 2010;29:3245
Pocock et al. Eur Heart J 2012;33:176

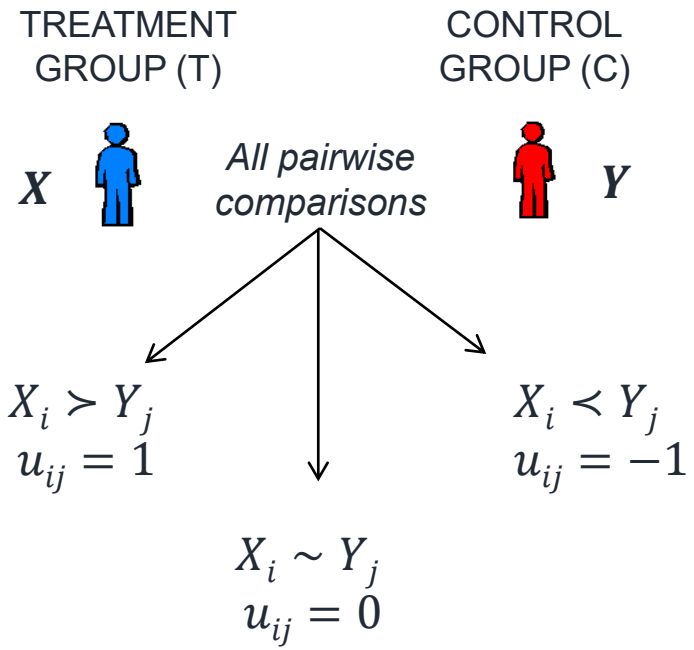
GPC – time to event



- Denote X_i^+ and Y_j^+ censored observations
- $u_{ij} = \begin{cases} +1 \; if \; X_i > Y_j \; or \; X_i^+ \geq Y_j \\ -1 \; if \; X_i < Y_j \; or X_i \leq Y_j^+ \\ 0 \; otherwise \end{cases}$

Gehan. Biometrika 1965;52:203

GPC – binary outcome

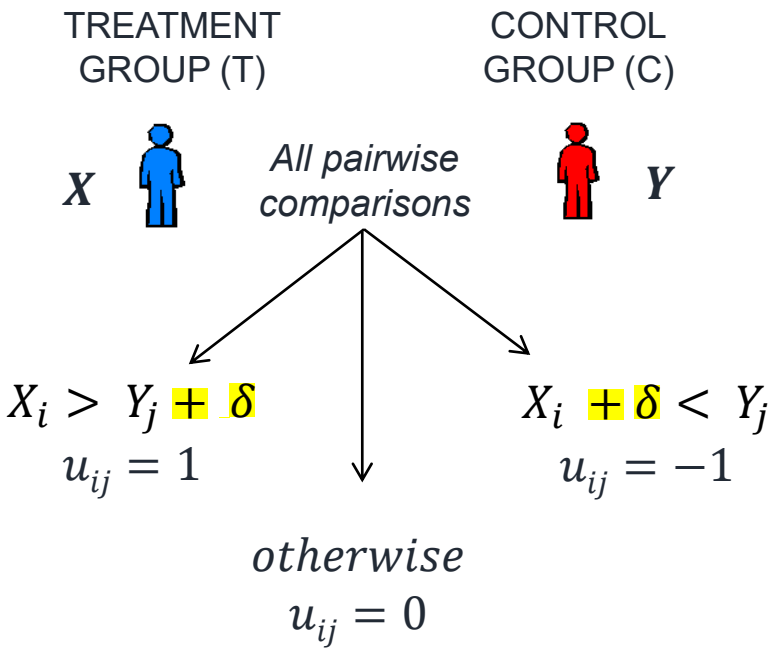


- Denote successes by 1 and failures by 0
- $u_{ij} = \begin{cases} +1 \; if \; X_i > Y_j \\ -1 \; if \; X_i < Y_j \\ 0 \; otherwise \end{cases}$

For binary outcomes, $U = P_T - P_C$,
the difference between the probabilities of success

Buyse. Stat Med 2010;29:3245

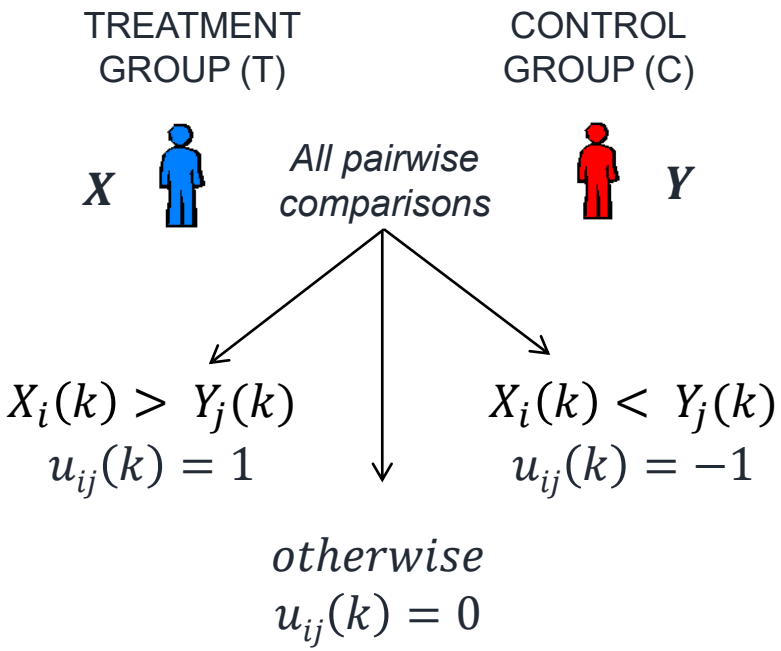
GPC – threshold of clinical similarity



1. Perform pairwise comparisons between all elements of ordered outcomes $\boldsymbol{X}$ and $\boldsymbol{Y}$
2. Calculate $u_{ij} = \begin{cases} +1 \; if \; X_i > Y_j + \delta \\ -1 \; if \; X_i + \delta < Y_j \\ 0 \; otherwise \end{cases}$
3. The statistic
   $U = \frac{1}{m \cdot n} \sum_{i=1}^{m} \sum_{j=1}^{n} u_{ij}$
   has a known distribution under H_0

Buyse. Stat Med 2010;29:3245

GPC – multiple weighted outcomes

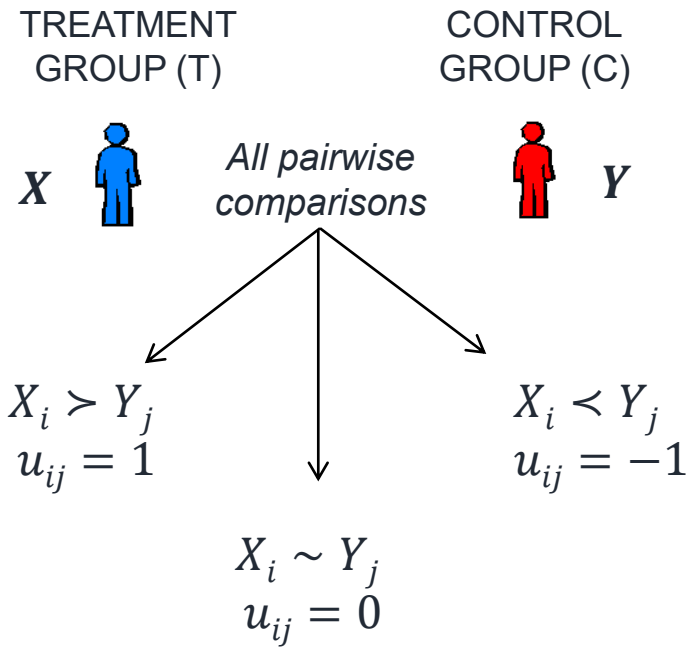


1. Perform pairwise comparisons between all elements of $\boldsymbol{X}$ and $\boldsymbol{Y}$
2. Calculate $u_{ij}(k) = \begin{cases} +1 \; if X_i(k) > Y_j(k) \\ -1 \; if X_i(k) < Y_j(k) \\ 0 \; otherwise \end{cases}$
3. The statistic $U = \frac{1}{m \cdot n} \sum_{k=1}^{K} \sum_{i=1}^{m} \sum_{j=1}^{n} w(k) u_{ij}(k)$ has a known distribution under H_0

Note: weights $w(k)$ are arbitrary, usually chosen so that $\sum_{k=1}^{K} w(k) = 1$

O'Brien. Biometrics 1984;69:1079

GPC – multiple prioritized outcomes

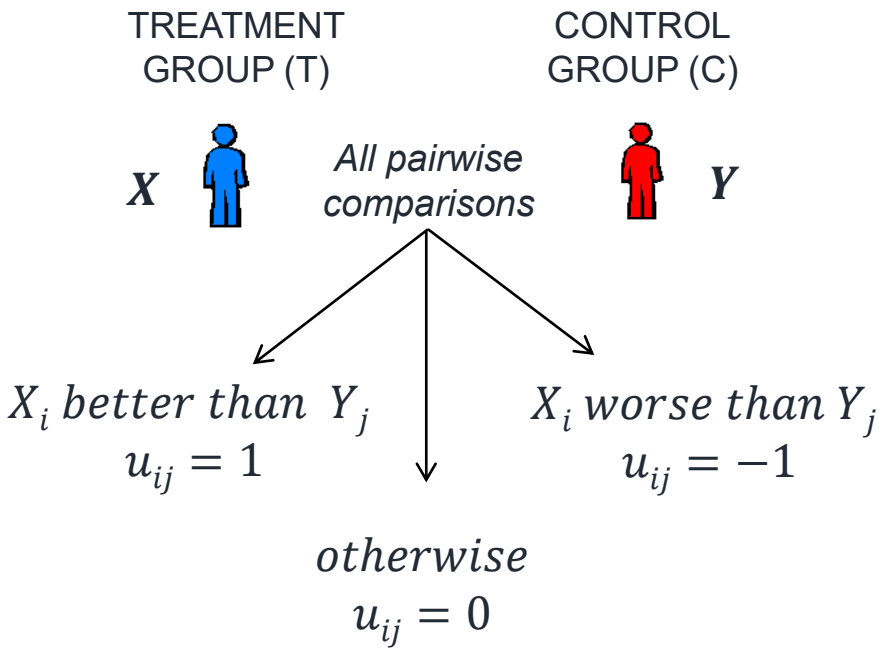


Outcome of 1st priority	Outcome of 2nd priority	Overall
Win	-	Win
Loss	-	Loss
Neutral	Win	Win
	Loss	Loss
	Neutral	Neutral

Note: priorities may be patient-centric

Buyse. Stat Med 2010;29:3245

Net Treatment Benefit (*NTB*)



The Net Treatment Benefit (*NTB*) is a *U*-statistic

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

$$= \frac{\#Wins - \#Losses}{\#Pairs}$$

Hoeffding. Ann Math Stat 1948;19:293

Measures of treatment effect

$$Finkelstein\text{-}Schoenfeld\ statistic^{1} = \#Wins - \#Losses$$

$$NTB^{2} = \frac{\#Wins - \#Losses}{\#Pairs}$$

$$Win\ Ratio^{3} = \frac{\#Wins}{\#Losses}$$

$$Win\ Odds^{4,5} = \frac{\#Wins + \frac{1}{2}\#(Neutral)}{\#Losses + \frac{1}{2}\#(Neutral)}$$

Note

$$NTB = \frac{Win\ Odds - 1}{Win\ Odds + 1}$$

[1] *Finkelstein & Schoenfeld. Stat Med 1999;18:1341*
[2] *Buyse. Stat Med 2010;29:3245*
[3] *Pocock et al. Eur Heart J 2012;33:176*
[4] *Dong et al. Stat Biopharm Res 2020;12:99*
[5] *Brunner et al. Stat Med 2021;40:3367*

NTB – interpretation

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

$$NTB = P(X > Y) - P(Y > X)$$

NTB is the *net* probability of a better outcome in one treatment group than in the other

More precisely, NTB is the probability that a patient taken at random in the treatment group has a better outcome than a patient taken at random in the control group, minus the probability of the opposite situation.

Péron et al. J Amer Med Ass 2016;2:905

NTB – interpretation

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

$$NTB = P(X > Y) - P(Y > X)$$

NTB is the *net* probability of a better outcome in one treatment group than in the other

More precisely, NTB is the probability that a patient taken at random in the treatment group has a better outcome than a patient taken at random in the control group, minus the probability of the opposite situation.

NTB is *not* the the difference between the probability for a patient to have a better outcome in the Experimental group than in the Control group! This would be an individual causal treatment effect. NTB is an average (population-level) treatment effect.

Fay et al. Stat Med 2018;37:2991

NTB and probabilistic index

NTB is a linear transformation of the probabilistic index PI

$$NTB = 2 \cdot PI - 1$$

where

$$PI = P(X > Y) + \frac{1}{2}P(X = Y)$$

PI ranges from 0 to 1, with ½ indicating no overall treatment effect

PI is closely related to the proportion of similar responses [1], the concordance index [2] the probability of overlap [3], and the area under the ROC curve [4].

[1] *Rom & Wang. Stat Med 1996;15:1489*
[2] *Harrell. Regression Model Strategies, 2001*
[3] *Stine & Heyse. Stat Med 2001;20:215*
[4] *Brumback et al. Stat Med 2006;25:575*

Inference and software for GPC

https://osf.io/xtd89/

R-package: BuyseTest

R code

```
> install.packages("BuyseTest", quiet = TRUE)
> library(BuyseTest)
```

Current version 3.0.2

```
BuyseTest(trt~tte(time, status = "status"), data = data)

BuyseTest.options()
```

Default: 2-sided test of a prioritized GPC, using first order U-statistic inference, with “Péron” scoring rule and expressing the treatment effect as the NTB

Inference with GPC (method.inference="…")

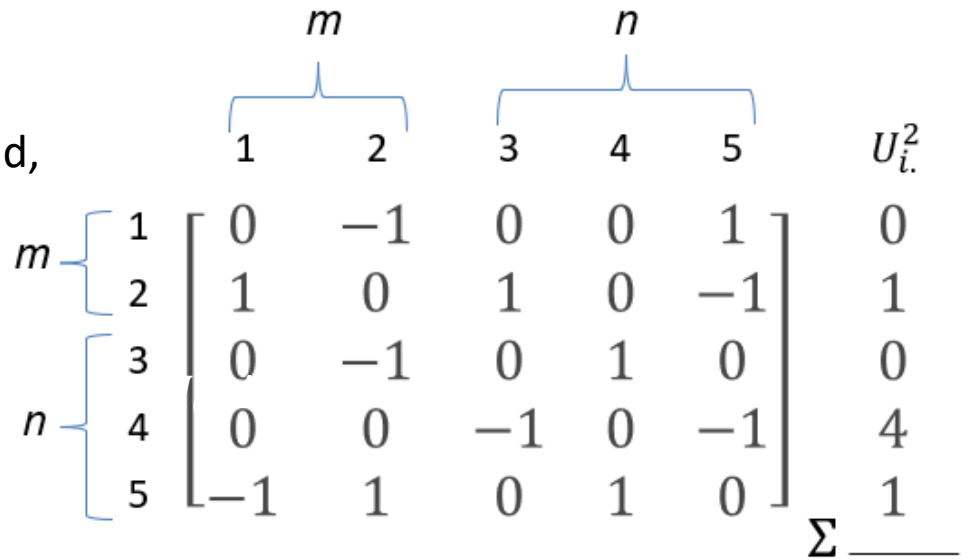
permutation

$H_0: F_E = F_C$

- Gehan, Finkelstein-Schoenfeld, exact permutation
- Type I error control in small samples
- No CI

studentized permutation

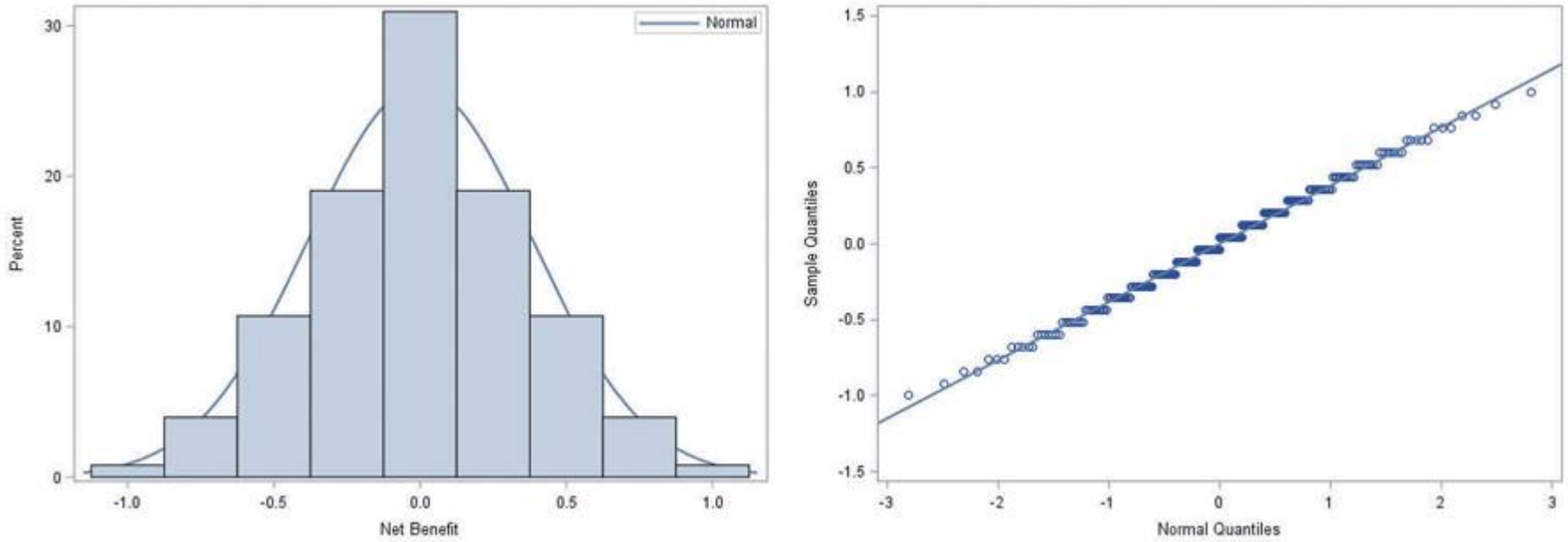
$H_0: \Delta = 0$



$$\hat{\sigma}_{\Delta}^2 = \frac{1}{N(N-1)mn}\sum_{i=1}^{N} U_{i.}^2$$

Verbeeck et al. J Biopharm Stat. (2020)
Anderson et al. Mathematics (2023)

Small sample inference with GPC: permutation



Histogram with fitted normal density curve (left) and normal Q-Q plot (right) of the exact permutation distribution of the net treatment benefit for a simulation of **five subjects per arm**.

Inference with GPC (method.inference="…")

u-statistic

$H_0: \Delta = 0$

- Hoefding decomposition
- Asymptotic method >n=m=30
- CI with transformation

$$\widehat{\sigma}^2_{uv} = \frac{N}{mn}\left((n-1)\delta^{uv}_{0,1} + (m-1)\delta^{uv}_{1,0} + \delta^{uv}_{1,1}\right)$$

bootstrap

$H_0: \Delta = 0$

Second H-decomposition = exact bootstrap (≈ ranked-based)

studentized bootstrap

$H_0: \Delta = 0$

- Similar performance as U-statistic
- But slower due to re-sampling

Verbeeck et al. J Biopharm Stat. (2020)
Anderson et al. Mathematics (2023)

Inference with GPC (method.inference="...")

- For n=m>30

  U-statistic/exact bootstrap – using the default inverse hyperbolic tangent transformation $x \rightarrow \frac{1}{2} log\left(\frac{1-x}{1+x}\right)$ for the NTB to be range- preserving

```
BuyseTest.options(order.Hprojection=2)

BuyseTest(trt~tte(time, status = "status"),method.inference="u-statistic", data = data)
```

- For n=m<=30

  (studentized) permutation for hypothesis testing

  CI?

General GPC : single outcome

binary (`b`, `bin`, or `binary`)

```
BuyseTest(treatment~bin(toxicity, operator = "<0"), data = data)
```

continuous (`c`, `cont`, or `continuous`)

```
BuyseTest(treatment~cont(score, operator = ">0", threshold=2), data = data)
```

time to event (`t`, `tte`, or `timetoevent`)

```
BuyseTest(treatment~tte(eventtime, status = "status", threshold=7), data = data)
```

```
          Generalized Pairwise Comparisons

Settings
   - 2 groups  : Control = C and Treatment = T
   - 1 endpoint:
      priority endpoint  type    operator
      1        toxicity  binary  lower is favorable

Point estimation and calculation of the iid decomposition

Estimation of the estimator's distribution
   - method: moments of the U-statistic

Gather the results in a S4BuyseTest object

 endpoint Delta
 toxicity  0.01
```

General GPC : multivariate outcomes

```
BT <- BuyseTest(treatment~tte(eventtime, status = "status")+bin(toxicity, operator = "<0"),
                scoring.rule="Gehan",data = data, trace=0)
```

S4-object: coef, confint, plot, print, summary to extract outputs

```
> summary(BT)
      Generalized pairwise comparisons with 2 prioritized endpoints

 - statistic       : net benefit (delta: endpoint specific, Delta: global)
 - null hypothesis : Delta == 0
 - confidence level: 0.95
 - inference       : H-projection of order 2 after atanh transformation
 - treatment groups: T (treatment) vs. C (control)
 - censored pairs  : deterministic score or uninformative
 - neutral pairs   : re-analyzed using lower priority endpoints
 - uninformative pairs: no contribution at the current endpoint, analyzed at later endpoints
 - results
  endpoint total(%) favorable(%) unfavorable(%) neutral(%) uninf(%)   delta   Delta CI [2.5% ; 97.5%]  p.value
eventtime   100.00        25.49          40.97       0.00    33.54 -0.1548 -0.1548 [-0.2785;-0.0261] 0.018620
 toxicity    33.54         9.31           7.62      16.61     0.00  0.0169 -0.1379  [-0.2761;0.0059] 0.060064
```

General GPC : output

More concise output:

```
> print(BT)
  endpoint total(%) favorable(%) unfavorable(%) neutral(%) uninf(%)   delta   Delta CI [2.5% ; 97.5%]  p.value
 eventtime   100.00        25.49          40.97       0.00    33.54 -0.1548 -0.1548 [-0.2785;-0.0261] 0.018620
  toxicity    33.54         9.31           7.62      16.61     0.00  0.0169 -0.1379  [-0.2761;0.0059] 0.060064
```

```
> confint(BT)
           estimate         se   lower.ci     upper.ci null    p.value
eventtime  -0.1548 0.06473057 -0.2784850 -0.026064275    0 0.01861953
toxicity   -0.1379 0.07240503 -0.2760918  0.005877914    0 0.06006369
```

Number of pairs:

```
> print(BT, percentage=FALSE)
  endpoint total favorable unfavorable neutral uninf   delta   Delta CI [2.5% ; 97.5%]  p.value
 eventtime 10000      2549        4097       0  3354 -0.1548 -0.1548 [-0.2785;-0.0261] 0.018620
  toxicity  3354       931         762    1661     0  0.0169 -0.1379  [-0.2761;0.0059] 0.060064
```

General GPC : output

Win ratio

```
> confint(BT, statistic="winRatio")
           estimate        se  lower.ci  upper.ci null    p.value
eventtime 0.6221626 0.1257302 0.4186857 0.9245271    1 0.01886039
toxicity  0.7161967 0.1268461 0.5061455 1.0134195    1 0.05947054
```

Probabilistic index and Success Odds

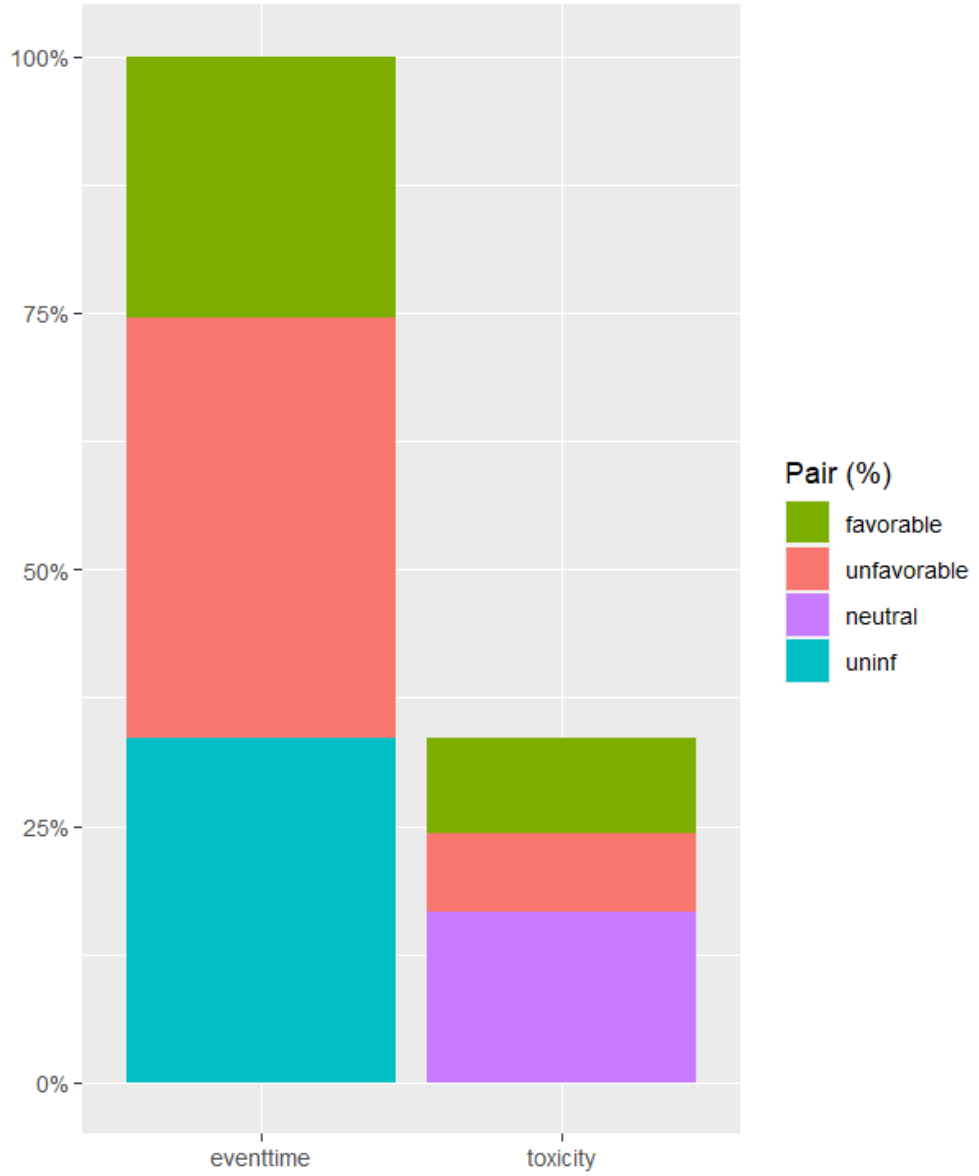
```
> BT <- BuyseTest(treatment~tte(eventtime, status = "status")+bin(toxicity, operator = "<0"),
+       scoring.rule="Gehan",data = data, trace=0, add.halfNeutral = TRUE)
> confint(BT, statistic="favorable")
           estimate        se  lower.ci  upper.ci null      p.value
eventtime   0.25490 0.03428485 0.1936518 0.3276484  0.5 2.814160e-09
toxicity    0.43105 0.03625982 0.3618483 0.5030534  0.5 6.046957e-02
> confint(BT, statistic="winRatio")
           estimate        se  lower.ci  upper.ci null    p.value
eventtime 0.6221626 0.1257302 0.4186857 0.9245271    1 0.01886039
toxicity  0.7576237 0.1120153 0.5670256 1.0122888    1 0.06046957
```

General GPC : scoring matrix

```
> BT <- BuyseTest(treatment~tte(eventtime, status = "status")+bin(toxicity, operator = "<0"),
+                 scoring.rule="Gehan",data = data, trace=0, keep.pairScore = TRUE)
> getPairScore(BT)
$eventtime
       index.C index.T favorable unfavorable neutral uninf weight
    1:       1     101         0           1       0     0      1
    2:       2     101         0           1       0     0      1
    3:       3     101         0           1       0     0      1
    4:       4     101         0           1       0     0      1
    5:       5     101         0           1       0     0      1
   ---
 9996:      96     200         0           0       0     1      1
 9997:      97     200         0           0       0     1      1
 9998:      98     200         0           0       0     1      1
 9999:      99     200         1           0       0     0      1
10000:     100     200         1           0       0     0      1

$toxicity
      index.C index.T favorable unfavorable neutral uninf weight
   1:      18     101         1           0       0     0      1
   2:      19     101         0           0       1     0      1
   3:      60     101         0           0       1     0      1
   4:       5     102         0           0       1     0      1
   5:       8     102         0           0       1     0      1
  ---
3350:      92     200         0           0       1     0      1
3351:      94     200         1           0       0     0      1
3352:      96     200         0           0       1     0      1
3353:      97     200         1           0       0     0      1
3354:      98     200         1           0       0     0      1
```

General GPC :Graphical display



Examples revisited

1. Time-to-first revisited: time-to-worst event

Or at least a simulated dataset resembling CHARM preserved (CHARM_sim.csv)

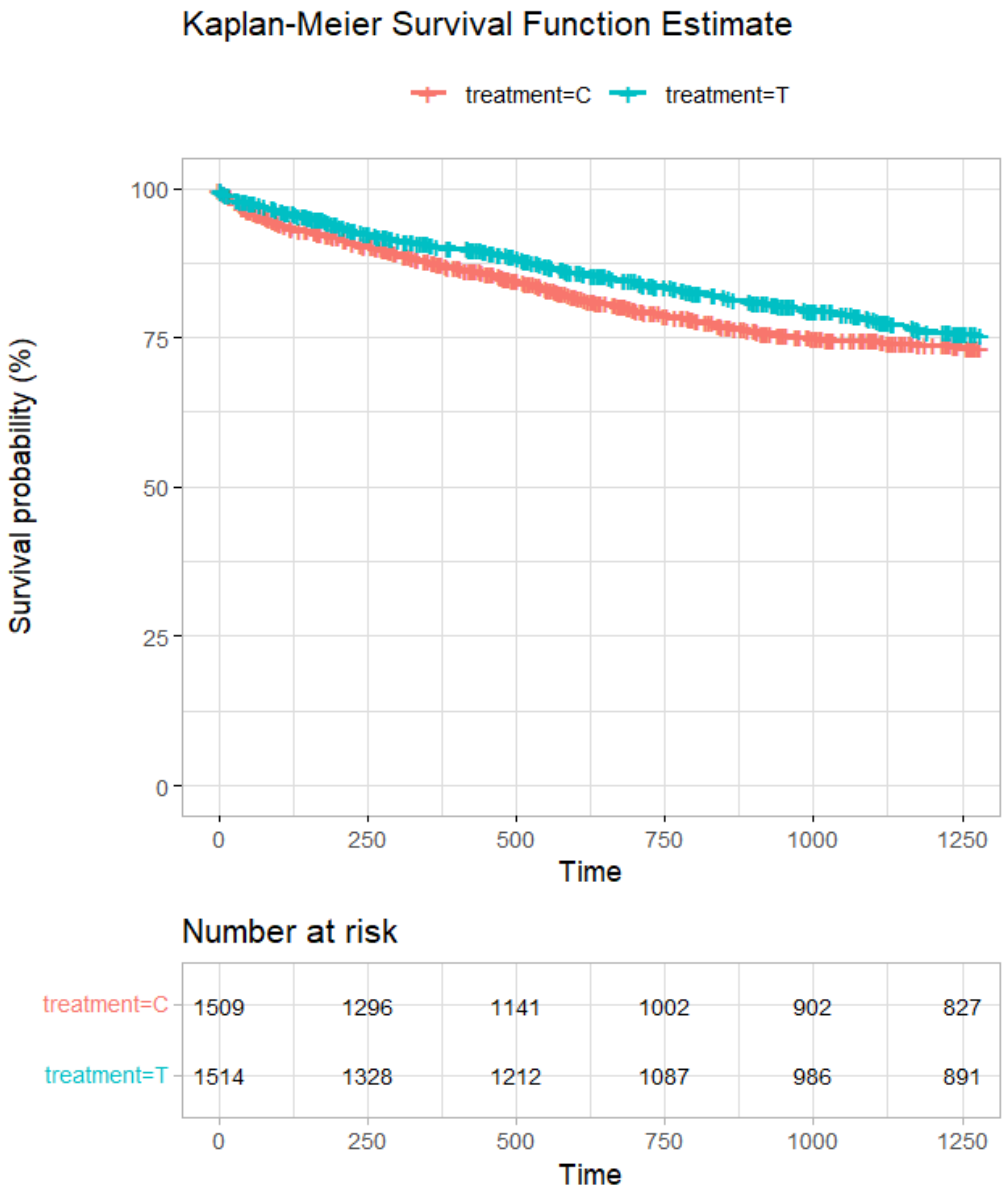
```
> head(charm)
  treatment Mortality statusMortality Hospitalization statusHospitalization Composite statusComposite
1         C  3.891039               1        3.891039                     0  3.891039               1
2         C  3.929701               1        3.929701                     0  3.929701               1
3         C  7.115158               1        7.115158                     0  7.115158               1
4         C  7.691922               1        7.691922                     0  7.691922               1
5         C 12.638044               1       12.638044                     0 12.638044               1
6         C 12.976806               1       12.976806                     0 12.976806               1
```

	Candesartan (n=1514)	Placebo (n=1509)
Cardiovascular death or hospital admission for CHF	333 (22·0%)	366 (24·3%)
Cardiovascular death	170 (11·2%)	170 (11·3%)
Hospital admission for CHF	241 (15·9%)	276 (18·3%)

Events in time-to-first composite

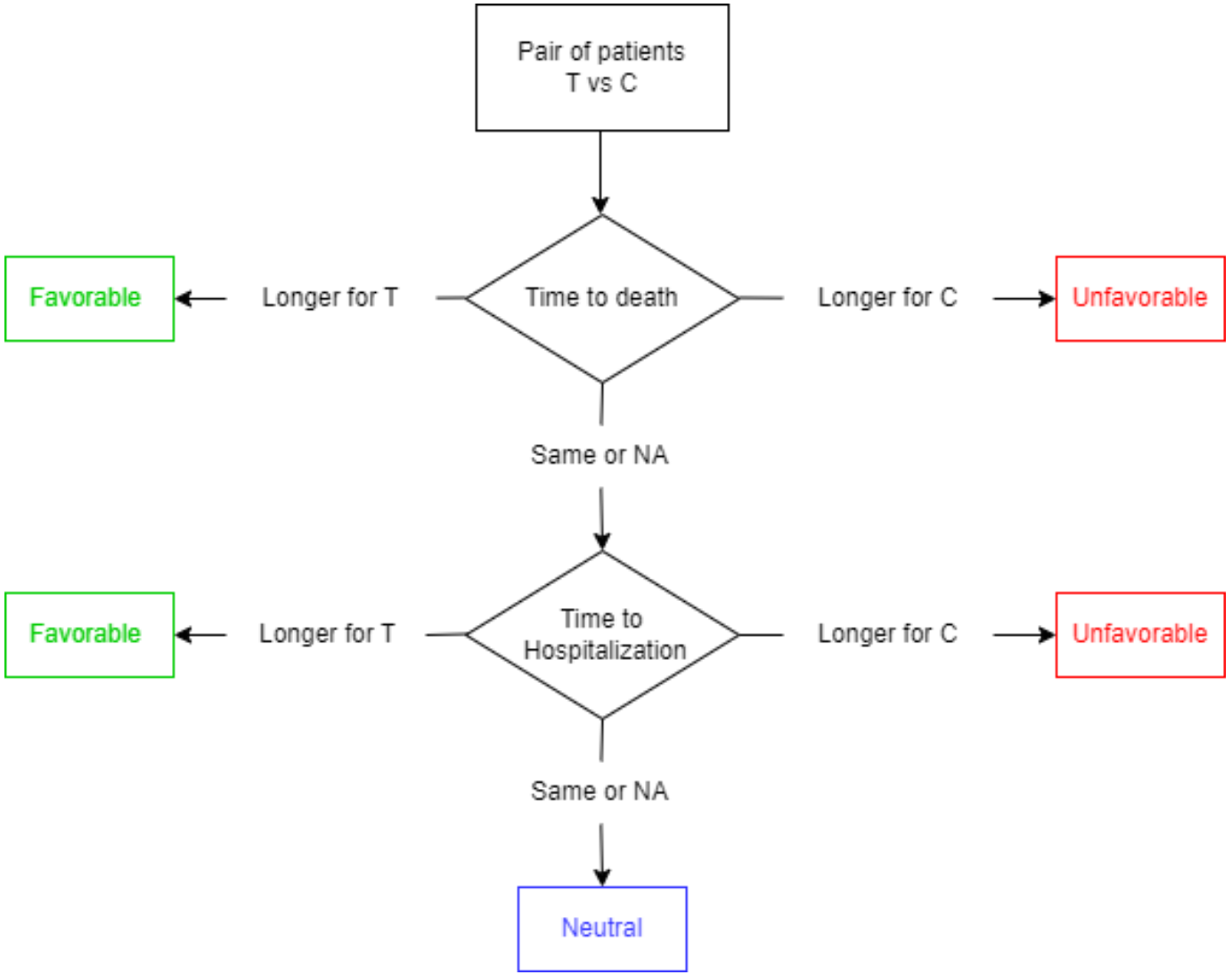
Candesartan	Placebo
92 (54%)	90 (53%)
241 (100%)	276 (100%)

Time-to-first *vs.* time-to-worst event



HR: 0.87 [0.75;1.01]
p-value: 0.0586

Time-to-first *vs.* time-to-worst event



Time-to-first *vs.* time-to-worst event

```
> BT_charm <- BuyseTest(treatment~tte(Mortality,statusMortality) + tte (Hospitalization,statusHospitalization),
+                       data=charm, scoring.rule = "Gehan", trace=0)
> summary(BT_charm)
       Generalized pairwise comparisons with 2 prioritized endpoints

 - statistic       : net benefit (delta: endpoint specific, Delta: global)
 - null hypothesis : Delta == 0
 - confidence level: 0.95
 - inference       : H-projection of order 2 after atanh transformation
 - treatment groups: T (treatment) vs. C (control)
 - censored pairs  : deterministic score or uninformative
 - uninformative pairs: no contribution at the current endpoint, analyzed at later endpoints
 - results
        endpoint total(%) favorable(%) unfavorable(%) neutral(%) uninf(%)  delta  Delta CI [2.5% ; 97.5%]  p.value
      Mortality   100.00         9.51           9.08          0    81.41 0.0042 0.0042  [-0.0157;0.0241] 0.676327
 Hospitalization    81.41        10.58           7.94          0    62.90 0.0264 0.0306   [0.0029;0.0582] 0.030108 *
```

Time-to-first *vs.* time-to-worst event (with threshold)

```
> BT_charm <- BuyseTest(treatment~tte(Mortality,statusMortality, threshold=14) +
+             tte (Hospitalization,statusHospitalization, threshold=14), data=charm, scoring.rule = "Gehan", trace=0)
> summary(BT_charm)
       Generalized pairwise comparisons with 2 prioritized endpoints

 - statistic       : net benefit (delta: endpoint specific, Delta: global)
 - null hypothesis : Delta == 0
 - confidence level: 0.95
 - inference       : H-projection of order 2 after atanh transformation
 - treatment groups: T (treatment) vs. C (control)
 - censored pairs  : deterministic score or uninformative
 - neutral pairs   : re-analyzed using lower priority endpoints
 - uninformative pairs: no contribution at the current endpoint, analyzed at later endpoints
 - results
        endpoint threshold total(%) favorable(%) unfavorable(%) neutral(%) uninf(%)  delta  Delta CI [2.5% ; 97.5%]
      Mortality        14   100.00         9.47           8.95       0.03    81.55 0.0052 0.0052   [-0.0147;0.025]
Hospitalization        14    81.58        10.51           7.94       0.04    63.09 0.0257 0.0308   [0.0033;0.0584]
 p.value
0.609892
0.028366 *
```

2. Inherently multivariate outcome revisited

- GPC analysis takes multiple prioritized outcomes into account:
  1. WHO grade 4 SOM
  2. WHO grade 3 SOM
  3. Number of days of SOM (shorter better, threshold of 1 week)
  4. Number of days to SOM (later better, threshold of week)

- Such an analysis is more clinical relevant *and* more powerful

GPC analysis of multivariate outcome

Prioritized outcomes	Favorable pairs	Favor Treatment	Neutral	Favor Control	Contrib	NTB
Grade 4 SOM	Yes vs No	26%	58%	16%	10%	10% (*P*=0.044)
Grade 3 SOM	Yes vs No	14%	33%	11%	3%	13% (*P*=0.022)
SOM duration	Longer by ≥ 1 week	10%	17%	6%	4%	17% (*P*=0.0033)
Time to SOM	Later by ≥ 1 week	4%	11%	2%	2%	19% (*P*=0.0012)
Overall		54%		35%		19% (*P*=0.0012)

GPC analysis of multivariate outcome

- NTB = 54% - 35% = 19%
  - NTB has straightforward interpretation
  - NNT (Number Needed to Treat) = 1/NTB
  - Contributions of all outcomes to NTB are additive
- Win Ratio = 54% / 35% = 1.54
  - Interpretation of Win Ratio challenging (except for single outcome under proportional hazards)

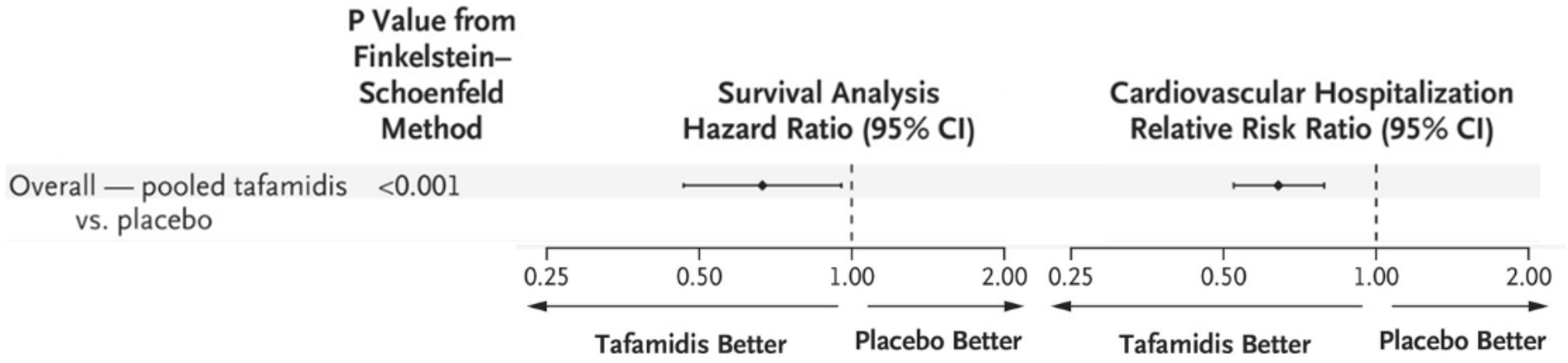
The NEW ENGLAND JOURNAL of MEDICINE

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 ?

What is lacking in this analysis?

Recall generalized U-statistic

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

Using superscripts W, L for Wins, Losses and subscripts $1, 2$ for outcomes:

$$\widehat{NTB} = U^W - U^L = (U_1^W + U_2^W) - (U_1^L + U_2^L) = (U_1^W - U_1^L) + (U_2^W - U_2^L)$$

NTB can be decomposed into two additive contributions :

$$\widehat{NTB} = \widehat{NTB_1} + \widehat{NTB_2}$$

Marginal NTB
for survival

Conditional NTB
for hospitalization

What is lacking in this analysis?

$$\widehat{WR} = \frac{U^W}{U^L} = \frac{U_1^W + U_2^W}{U_1^L + U_2^L}$$

$\widehat{WR}$ cannot be decomposed into the contributions of each outcome:

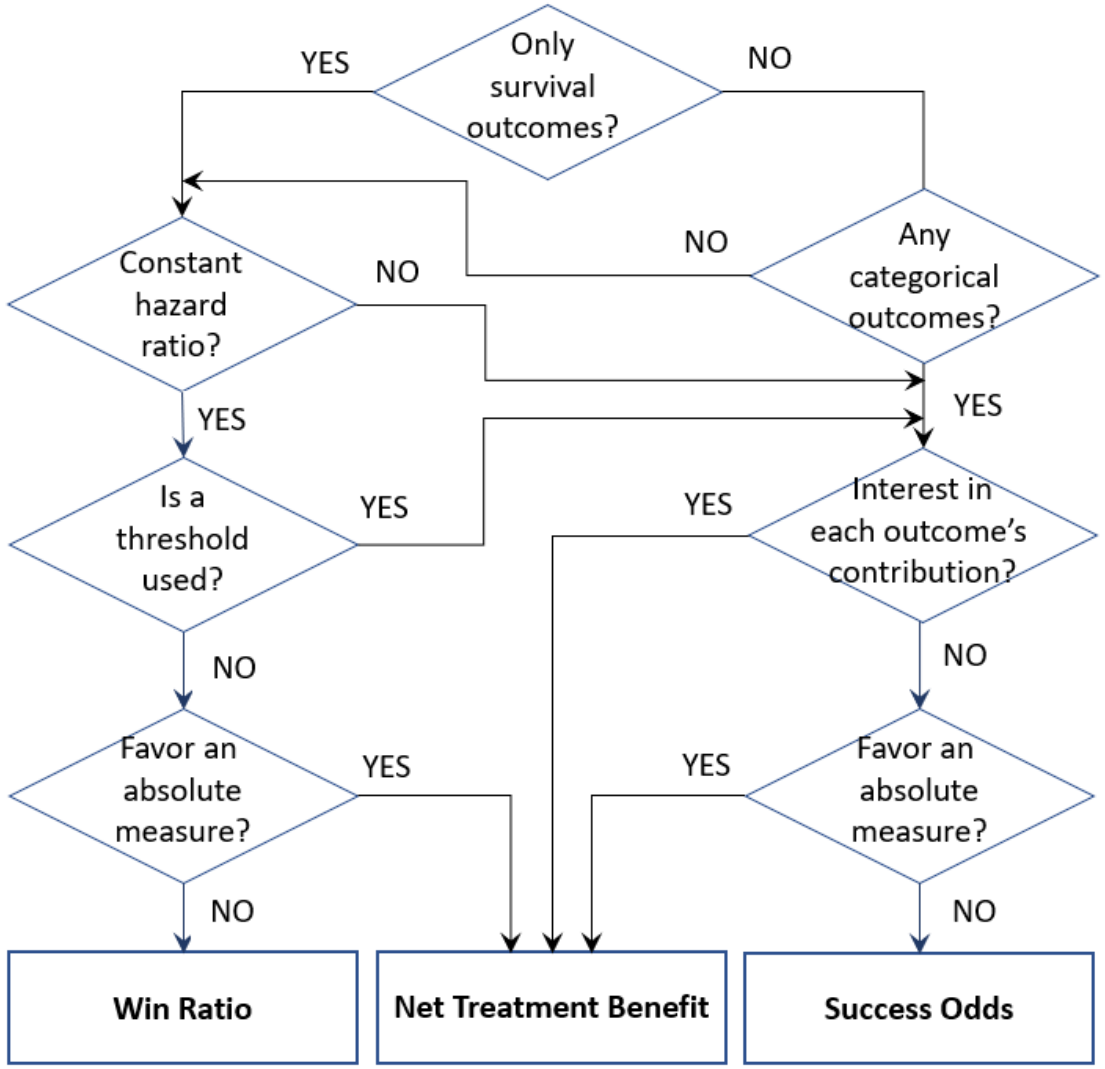
$$\widehat{WR_1} = \frac{U_1^W}{U_1^L}$$

Marginal WR for survival (reciprocal of hazard ratio)

$$\widehat{WR_2} = \frac{U_2^W}{U_2^L}$$

Conditional WR for hospitalization (odds ratio)

Guidance of GPC measures in clinical trials



Verbeeck et al. *JACC 2023*

3. EB rare disease trial revisited

- 16 pediatric subjects treated with placebo and diacerin cream in a longitudinal cross-over trial (14 paired)

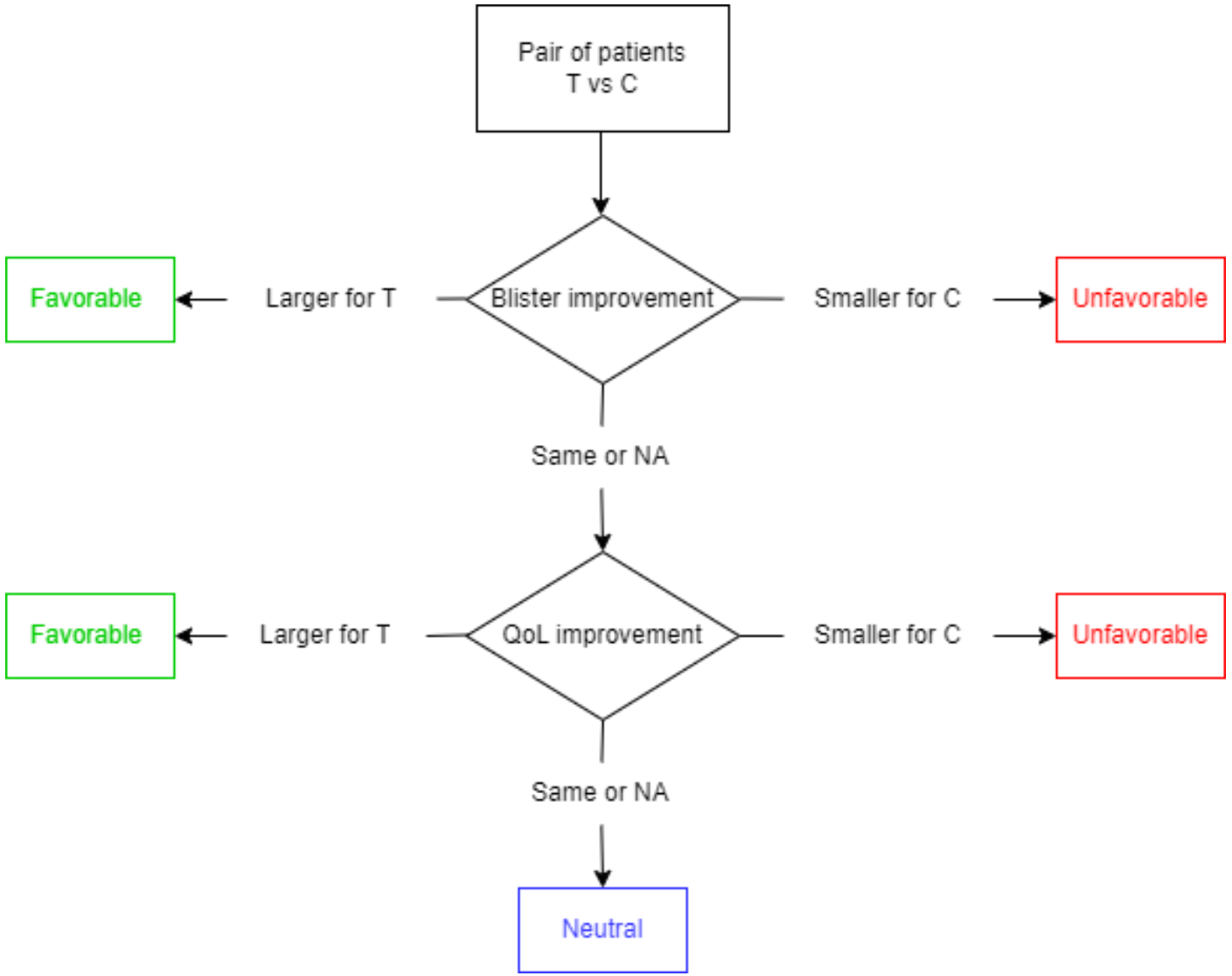
- Patient-centric analysis: blister count and change in QoL at week 4

- Uncertainty in blister counts

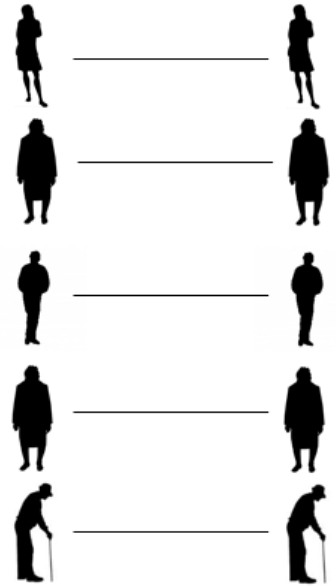
EB revisited:

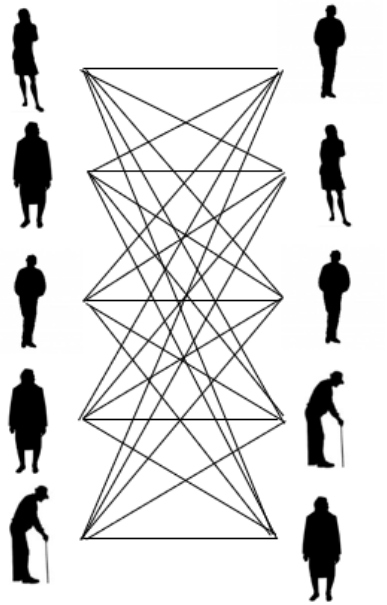
```
> head(EB)
    Id Time Group StdDiffCount Bin DiffQoL period
1 1001   t4     V    0.6666667   1       2      1
2 1001  t12     P    0.0000000   0       0      2
3 1002   t4     P   -0.2500000   0      -1      1
4 1002  t12     V   -4.0000000   0       0      2
5 1004   t4     V    0.5454545   1       1      1
6 1004  t12     P   -1.0000000   0       1      2
```

EB revisited: Patient-centric outcome



Matched versus unmatched GPC





$$\Delta_m = \mathbb{P}\left(Y_i^T > Y_i^C\right) - \mathbb{P}\left(Y_i^C > Y_i^T\right)$$

$$\Delta_{unm} = \mathbb{P}\left(Y_i^T > Y_j^C\right) - \mathbb{P}\left(Y_i^C > Y_j^T\right)$$

Conditional sign test : $Z_m = \frac{N_{YT} - N_{YC}}{\sqrt{N_{YT} + N_{YC}}} \sim N(0,1)$

requires at least 15-20 (paired) subjects

Matsouaka SMMR (2022)
Verbeeck et al. OJRD. (2023)
Konietschke et al. Electron J Stat (2012)

EB revisited:
Univariate insufficient evidence, but patient-centric analysis shows treatment effect

```
> print(BuyseTest(Group~b(Bin)+c(DiffQoL), data=EB,method.inference="permutation",n.resampling=10000))

         Generalized Pairwise Comparisons

Settings
   - 2 groups    : Control = P and Treatment = V
   - 2 endpoints:
       priority endpoint type        operator
       1        Bin      binary      higher is favorable
       2        DiffQoL  continuous  higher is favorable
   - neutral pairs: re-analyzed using lower priority endpoints

Point estimation

Estimation of the estimator's distribution
   - method: permutation test with 10000 permutations
   - cpus  : 1
  |++++++++++++++++++++++++++++++++++++++++++++++++++| 100% elapsed=04s

Gather the results in a S4BuyseTest object

 endpoint total(%) favorable(%) unfavorable(%) neutral(%) uninf(%)  delta  Delta   p.value
      Bin   100.00           44          10.67       45.33     0.00 0.3333 0.3333 0.1385861
  DiffQoL    45.33           32           6.22        5.33     1.78 0.2578 0.5911 0.0036996
```

EB revisited: less evidence for count outcome

```
> print(BuyseTest(Group~c(StdDiffCount)+c(DiffQoL), data=EB,method.inference="permutation",n.resampling=10000))

         Generalized Pairwise Comparisons

Settings
   - 2 groups   : Control = P and Treatment = V
   - 2 endpoints:
      priority endpoint      type        operator
      1        StdDiffCount  continuous  higher is favorable
      2        DiffQoL       continuous  higher is favorable
   - neutral pairs: re-analyzed using lower priority endpoints

Point estimation

Estimation of the estimator's distribution
   - method: permutation test with 10000 permutations
   - cpus  : 1
  |++++++++++++++++++++++++++++++++++++++++++++++++++| 100% elapsed=02s

Gather the results in a S4BuyseTest object

     endpoint total(%) favorable(%) unfavorable(%) neutral(%) uninf(%)  delta  Delta  p.value
 StdDiffCount   100.00        64.00          27.11       2.22     6.67 0.3689 0.3689 0.070993
      DiffQoL     8.89         6.22           0.44       1.78     0.44 0.0578 0.4267 0.041896
```

EB revisited: accounting for blister uncertainty

```
> print(BuyseTest(Group~c(StdDiffCount, threshold=0.2)+c(DiffQoL), data=EB,method.inference="permutation",n.resampling=1000
0))

         Generalized Pairwise Comparisons

Settings
   - 2 groups    : Control = P and Treatment = V
   - 2 endpoints:
       priority endpoint      type        operator             threshold
       1        StdDiffCount  continuous  higher is favorable  0.2
       2        DiffQoL       continuous  higher is favorable
   - neutral pairs: re-analyzed using lower priority endpoints

Point estimation

Estimation of the estimator's distribution
   - method: permutation test with 10000 permutations
   - cpus  : 1
  |++++++++++++++++++++++++++++++++++++++++++++++++++| 100% elapsed=02s

Gather the results in a S4BuyseTest object

     endpoint threshold total(%) favorable(%) unfavorable(%) neutral(%) uninf(%)  delta  Delta  p.value
 StdDiffCount       0.2      100        56.44          19.56      17.33     6.67 0.3689 0.3689 0.066093
     DiffQoL                 24        15.11           2.67       3.56     2.67 0.1244 0.4933 0.016998
```

EB revisited: CI consistent with p-value in this case

```
> print(BuyseTest(Group~b(Bin)+c(DiffQoL), data=EB,method.inference="u-statistic"), percentage=FALSE)

         Generalized Pairwise Comparisons

Settings
   - 2 groups    : Control = P and Treatment = V
   - 2 endpoints:
       priority endpoint type        operator
       1        Bin      binary      higher is favorable
       2        DiffQoL  continuous  higher is favorable
   - neutral pairs: re-analyzed using lower priority endpoints

Point estimation and calculation of the iid decomposition

Estimation of the estimator's distribution
   - method: moments of the U-statistic

Gather the results in a S4BuyseTest object

 endpoint total favorable unfavorable neutral uninf  delta  Delta CI [2.5% ; 97.5%]   p.value
      Bin   225        99          24     102     0 0.3333 0.3333  [-0.0291;0.6183] 0.0706270
  DiffQoL   102        72          14      12     4 0.2578 0.5911   [0.1931;0.8221] 0.0059238
```

permutation

```
  p.value
0.1406859
0.0036996
```

4. Benefit-Risk assessment revisited

- Simple situation (as on slide 19):
  - binary efficacy outcome (1 = response, 0 = no response)
  - binary safety outcome (1 = no toxicity, 0 = toxicity)

Outcomes	**Treatment**	**Control**	**Difference**
Response rate (benefit)	0.5	0.2	0.3
Toxicity rate (risk)	0.6	0	0.6
Marginal benefit-risk difference			-0.3

- GPC analyses of prioritized outcomes
  - Response > Toxicity
  - Toxicity > Response

4. Benefit-Risk assessment revisited

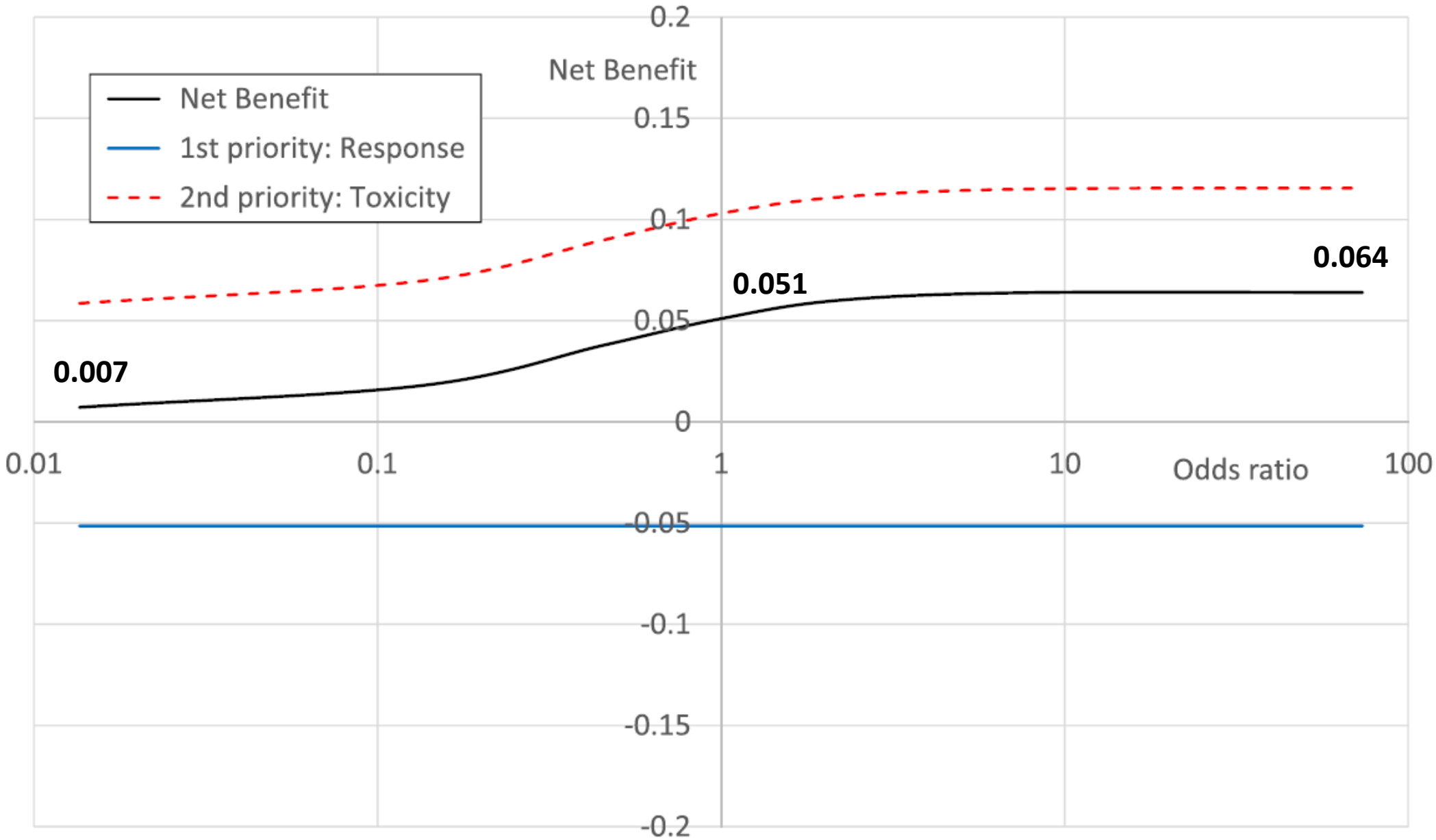
- Actual trial : « IDEA France » for patients with colorectal cancer (@4 years follow-up)
  - binary efficacy outcome (1 = « response » = Disease-Free , 0 = DFS event)
  - binary safety outcome (1 = no Grade 3 / 4 neurotoxicity, 0 = Grade 3 / 4 neurotoxicity)

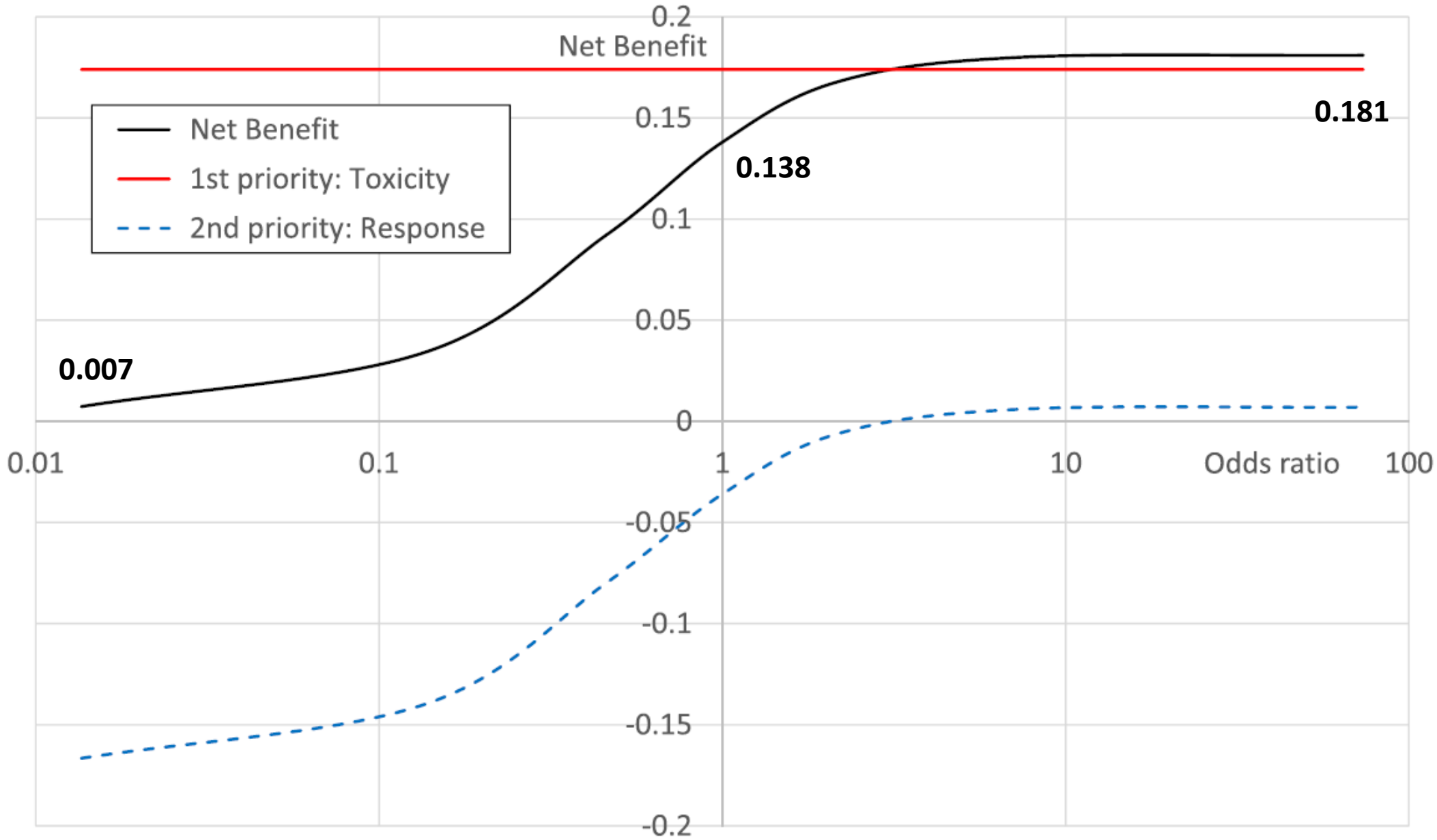
Outcomes	**3 months**	**6 months**	**Difference**
Disease-free (benefit)	0.69	0.74	- 0.05
Grade 3 / 4 neurotoxicity (risk)	0.08	0.26	- 0.18
Marginal benefit-risk difference			+ 0.13

- GPC analyses of prioritized outcomes
  - Response > Toxicity
  - Toxicity > Response

4. Benefit-Risk assessment revisited

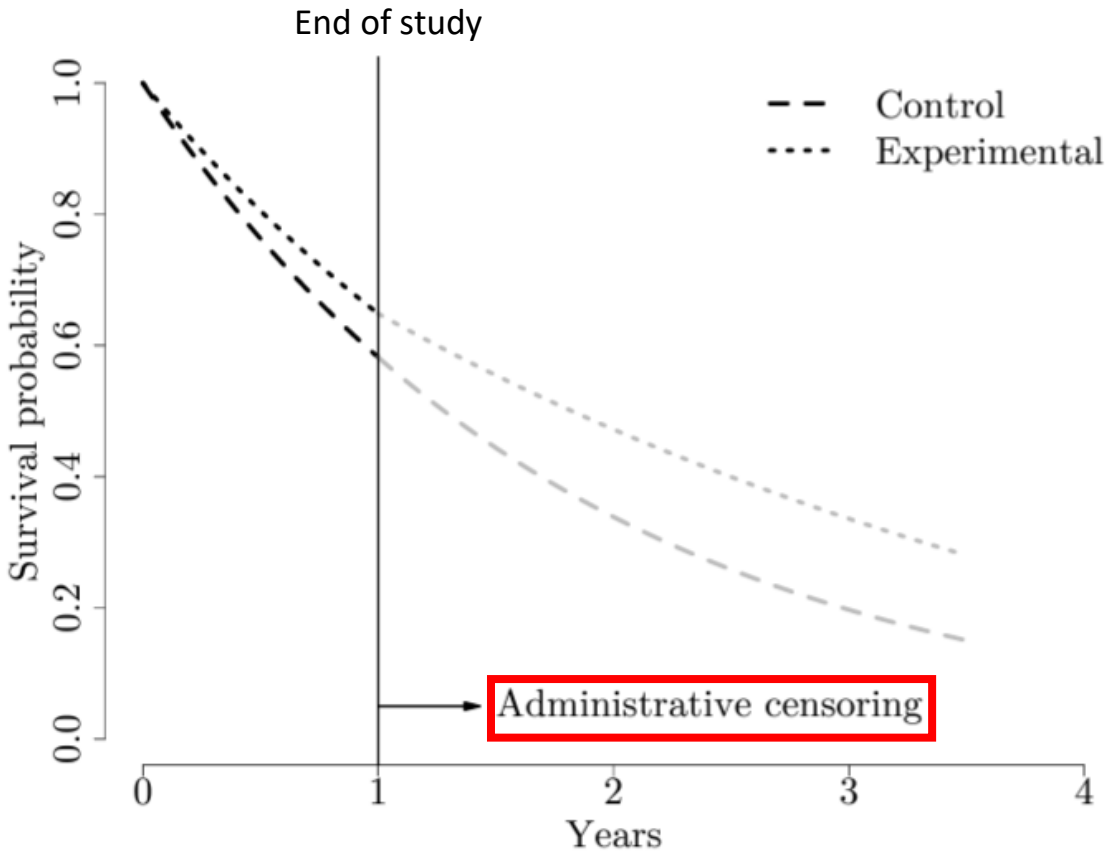
```
install.packages("BuyseTest")
library(BuyseTest)
# Load IDEA dataset under independence of DFS and GR_3_4_CNS
setwd("C:/Users/mbuyse/OneDrive - IDDI/Google Drive/data/BENEFIT/IDEA")
data <- read.delim(file="Ch_01_1_5_IDEA_indep.csv",sep=",")
# Experimental arm is 3-month treatment  (arm1=1), Control arm is 6-month treatment (arm1=0)
data$arm1 <- 1-data$arm
# GPC with DFS > GR_3_4_CNS
GPC <- BuyseTest(arm1 ~ b(Gr_3_4_PSN)+b(DFS),data = data)
summary(GPC)
# GPC with GR_3_4_CNS > DFS
GPC <- BuyseTest(arm1 ~ b(DFS)+b(Gr_3_4_PSN),data = data)
summary(GPC)
# Repeat with IDEA datasets under positive and negative association between DFS and GR_3_4_CNS
```

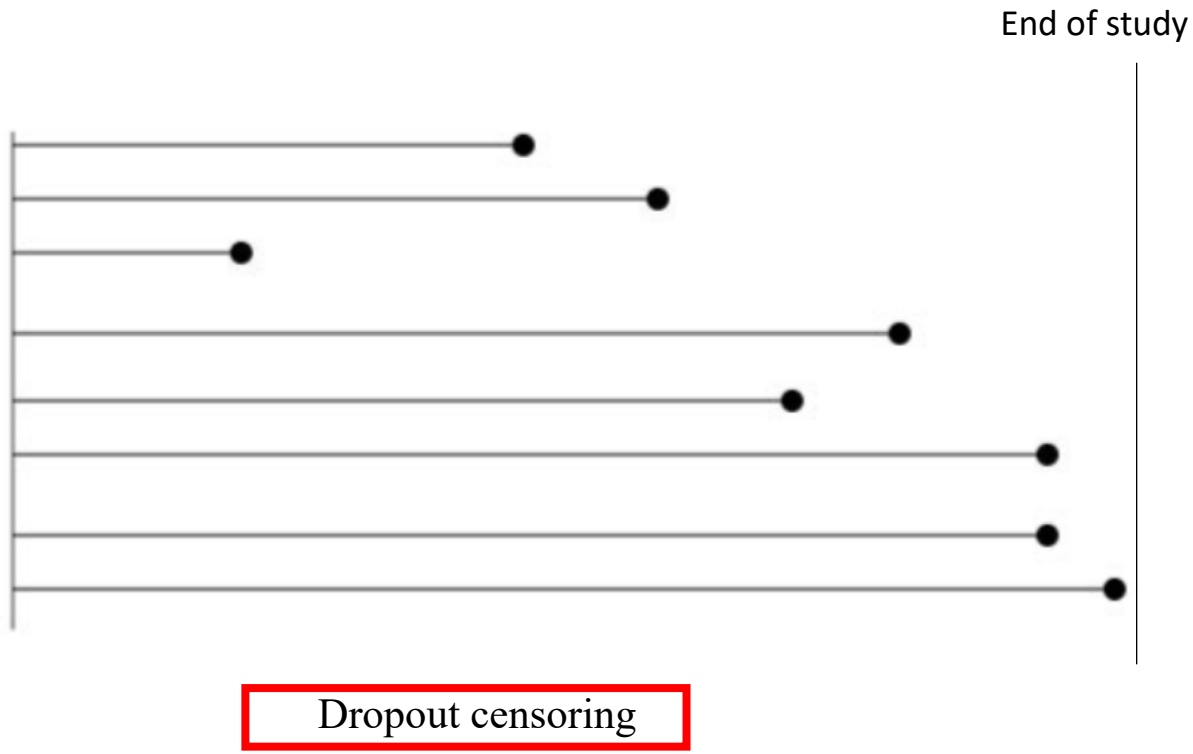

Net Benefit
1st priority: Response
2nd priority: Toxicity
Net Benefit
0.2
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
0.01
0.1
1
10
100
Odds ratio
0.007
0.051
0.064


Net Benefit
0.2
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
0.01
0.1
1
10
100
Odds ratio
Net Benefit
1st priority: Toxicity
2nd priority: Response
0.007
0.138
0.181

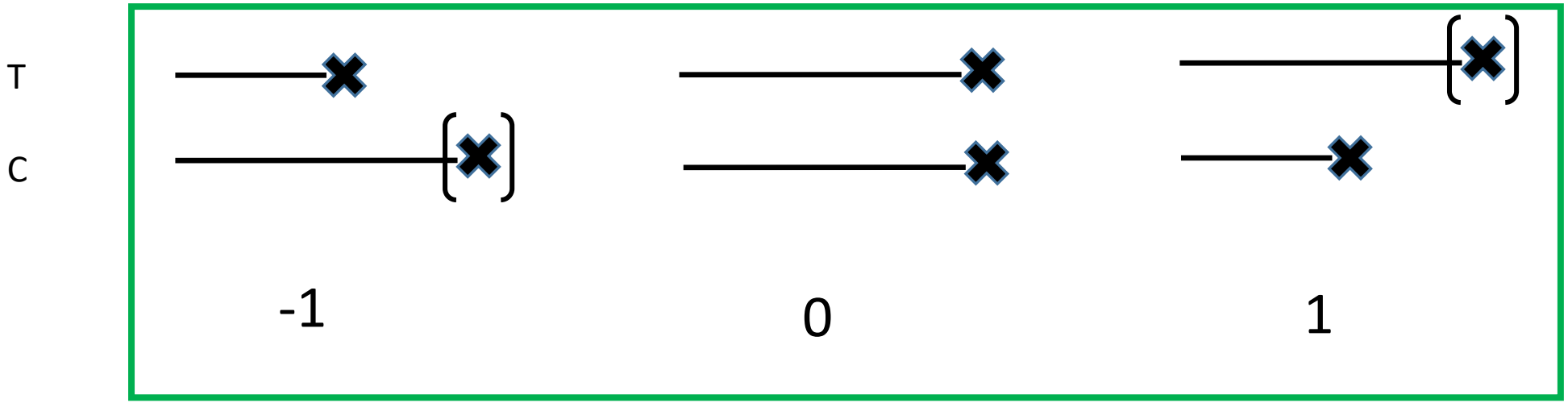
Advanced Topics

1. Censoring

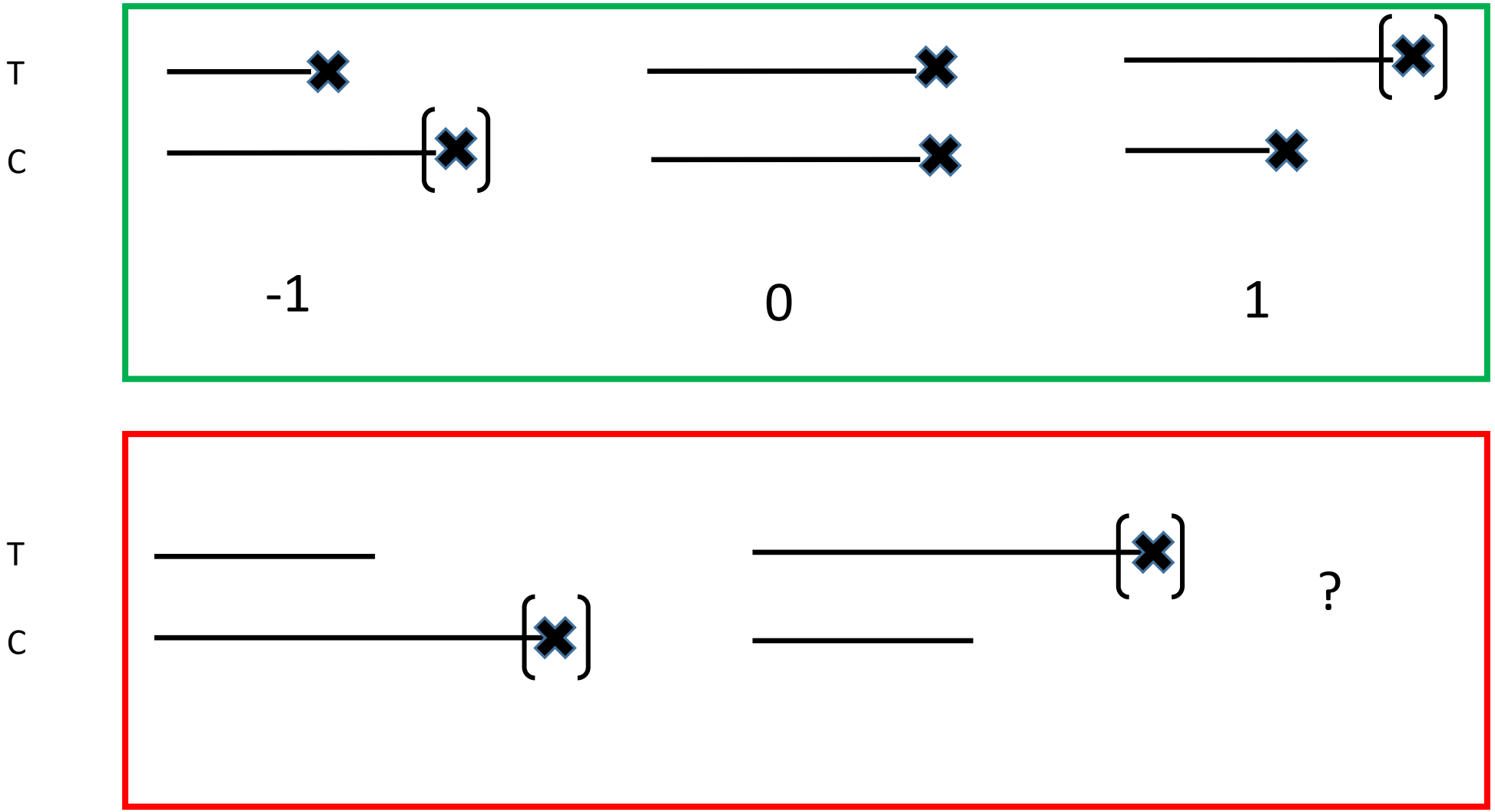

End of study
Control
Experimental
Survival probability
0.0
0.2
0.4
0.6
0.8
1.0
Administrative censoring
0
1
2
3
4
Years


End of study
Dropout censoring

Censoring induces informative



Censoring induces informative and uninformative pairs



Corrections for drop-out censoring

- Naïve approaches: (BuyseTest: scoring.rule = "Gehan“)
  - Gehan score: uninformative pairs = 0
  - Harrell score: ignore uninformative pairs, by considering only informative pairs

Under proportional hazards, the latter is unbiased

$$\mathrm{E}(\hat{\Delta}_c) = \frac{1}{P_{inf}} * \mathrm{E}(\hat{\Delta}_G) = \frac{\lambda_X + \lambda_Y + \lambda_{X^c} + \lambda_{Y^c}}{\lambda_Y + \lambda_X} * \frac{\lambda_Y - \lambda_X}{\lambda_X + \lambda_Y + \lambda_{X^c} + \lambda_{Y^c}} = \frac{\lambda_Y - \lambda_X}{\lambda_Y + \lambda_X} = \Delta.$$

Gehan Biometrika (1965)
Harrell et al. JAMA (1982)
Deltuvaite-Thomas et al. Biometrical journal (2022)

Corrections for drop-out censoring

- Imputation approaches: (BuyseTest: scoring.rule = “Peron”)
  - Use Kaplan-Meier estimates of survival function to estimate the probability of more favorable outcome per pair

    - Per treatment arm: Efron (Extended by Péron for unobserved last observation)

(δ_i^E, δ_i^C)	$Y_i^E > Y_j^C$	$Y_i^E = Y_j^C$	$Y_i^E < Y_j^C$
(1,1)	1	0	-1
(0,1)	1	1	$2\frac{\widehat{S_T^E}(Y_j^C)}{\widehat{S_T^E}(Y_i^E)} - 1$
(1,0)	$1 - 2\frac{\widehat{S_T^C}(Y_i^E)}{\widehat{S_T^C}(Y_j^C)}$	-1	-1
(0,0)	$1 - \frac{\widehat{S_T^C}(Y_i^E)}{\widehat{S_T^C}(Y_j^C)}$	0	$\frac{\widehat{S_T^E}(Y_j^C)}{\widehat{S_T^E}(Y_i^E)} - 1$

    - Joint distribution: Latta

Efron Proc 5th Berkeley Symposium on Mathematical Statistics and Probability (1967)
Péron et al. SMMR (2016)
Latta Biometrika (1977)
Deltuvaite-Thomas et al. Biometrical journal (2022)

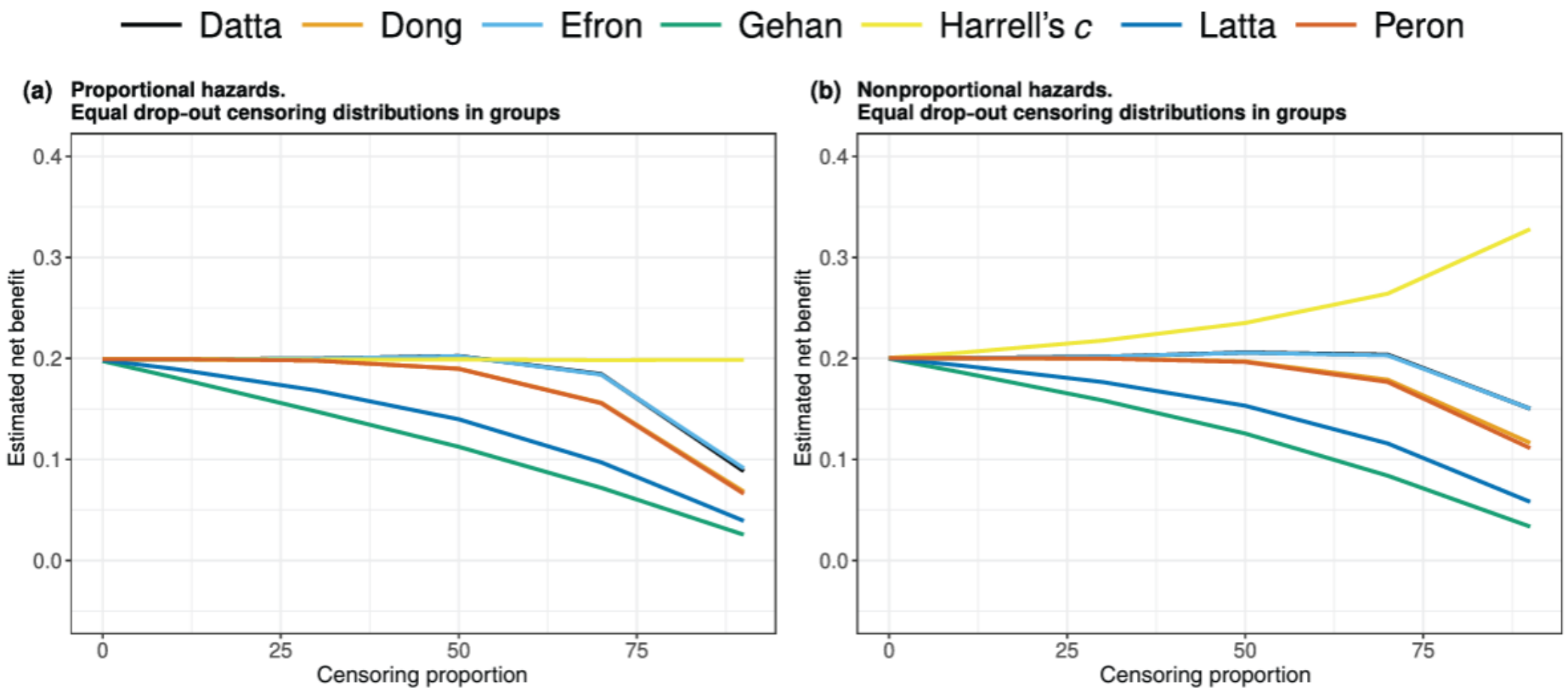
Corrections for drop-out censoring

- IPCW approaches: (BuyseTest: not implemented)
  - Weigh scores by the inverse of the probability of censoring, obtained by the Kaplan–Meier estimates of the censoring distribution.
    - Weighting a single score: Datta
    - Weighting a favorable and unfavorable score: Dong

IPCW by Datta = imputation by Efron

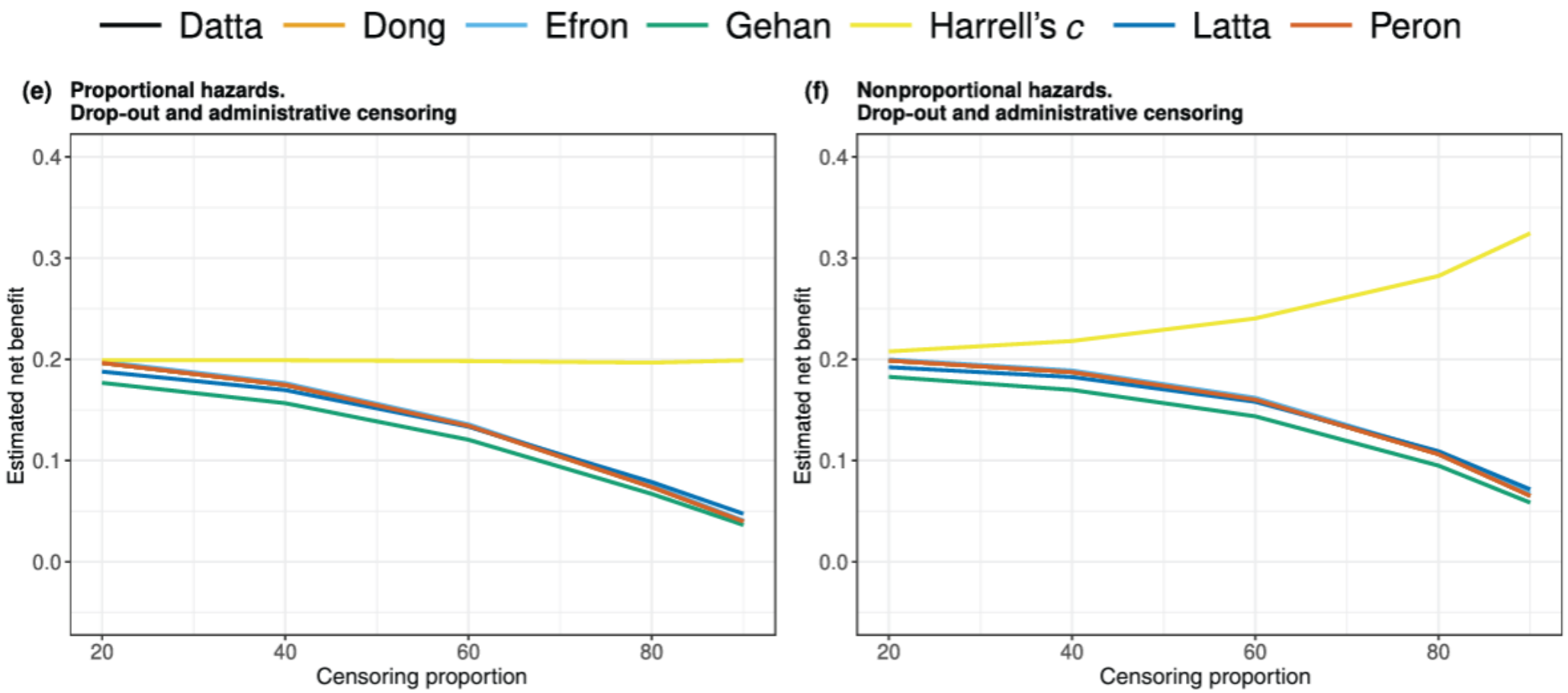
Datta et al. Scandinavian Journal of Statistics (2010)
Dong et al. Journal of Biopharmaceutical Statistics (2020)
Stute et al. Scandinavian Journal of Statistics (1994)
Deltuvaite-Thomas et al. Biometrical journal (2022)

Correction for drop-out censoring bias



Deltuvaite-Thomas et al. Biometrical journal (2022)

But not for administrative censoring bias



Deltuvaite-Thomas et al. Biometrical journal (2022)

Correction for administrative censoring: restricted NTB

~ Restricted mean survival time

The restricted NTB to time ϵ:

$$rNTB = \mathbb{P}\left(\min(Y_i^T, \epsilon) > Y_j^C\right) - \mathbb{P}\left(\min(Y_i^C, \epsilon) > Y_j^E\right)$$

Piffoux et al. J Clin Epidemiol (2024)

2. Stratification

- Recall the U-statistic

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

- With s strata, this formula naturally generalizes to

$$U = \sum_{k=1}^{s} \frac{1}{m_k \cdot n_k} \sum_{i=1}^{m_k} \sum_{j=1}^{n_k} u_{ijk}$$

- This U-statistic is a conditonal estimate of the NTB (given the strata)

Buyse. Stat Med 2010;29:3245

Choice of weights

- Unfortunately, this conditional estimate has undesirable properties
  - too much weight is given to large strata:
    e.g. two strata with 10 and 60 patients → weights of 25 and 900 if balanced
  - imbalances within strata inappropriately modify the weights:
    e.g. stratum of 10 patients → weight from 9 if imbalanced to 25 if balanced
- With s strata, a better estimate uses Cochran-Mantel-Haenszel weights

$$U = \sum_{k=1}^{s} \frac{m_k + n_k}{m + n} \sum_{i=1}^{m_k} \sum_{j=1}^{n_k} u_{ijk}$$

De Backer and Fay (to appear 2024)

Conditional vs. marginal estimates

- Conditional estimates have undesirable properties
  - they may differ from marginal estimates
  - they may have larger variance than marginal estimates
  - NTB and WR suffer from non collapsibility (as do odds ratios and hazard ratios)

Table 1: Non-collapsibility of the Odds Ratio in a Hypothetical Target Population

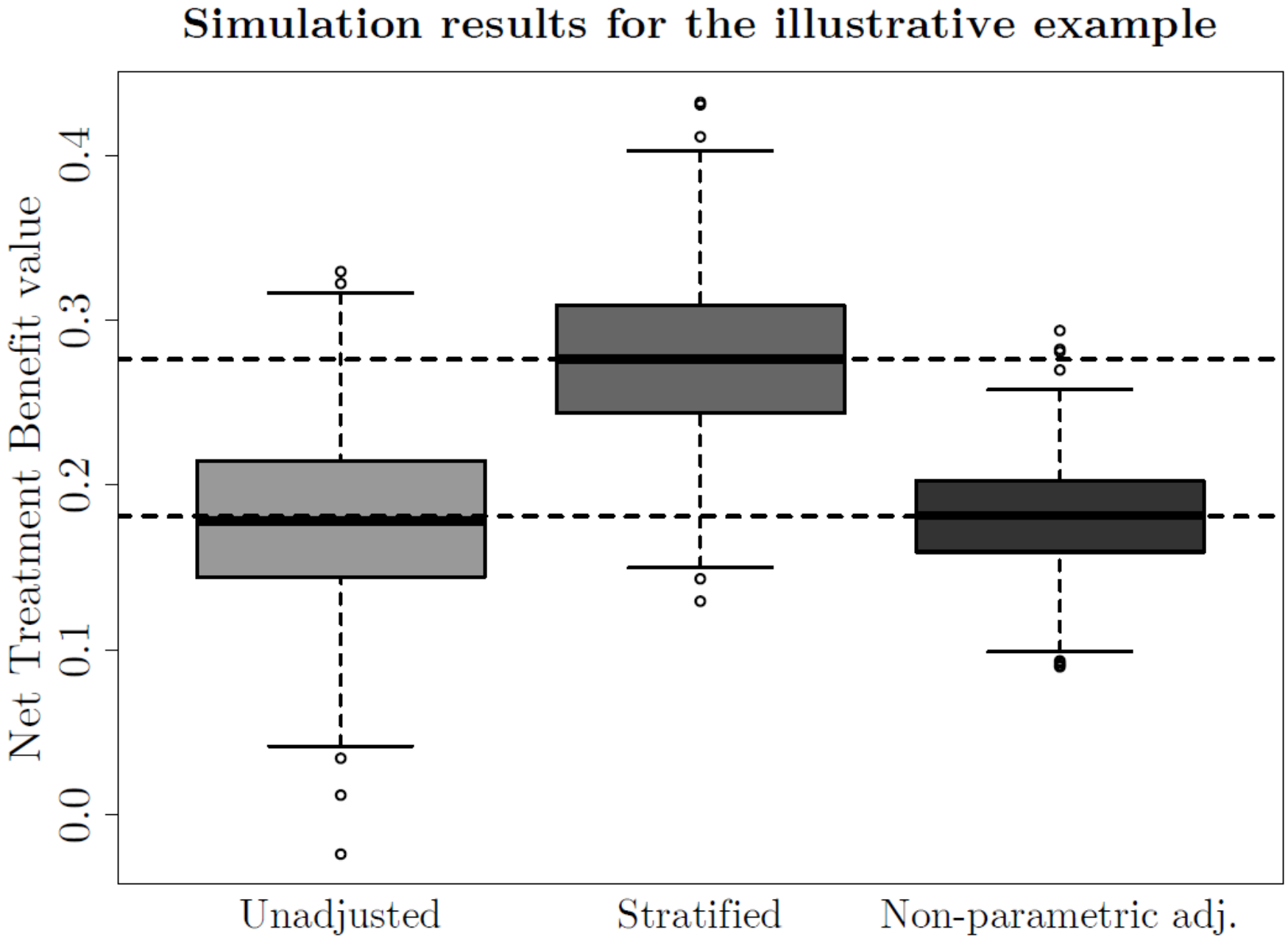
	Percentage of target population	Success rate		Odds ratio
		New drug	Placebo	
Biomarker-positive	50%	80.0%	33.3%	8.0
Biomarker-negative	50%	25.0%	4.0%	8.0
Combined	100%	52.5%	18.7%	4.8

De Backer and Fay (to appear 2024)

U.S. FOOD & DRUG ADMINISTRATION

Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products

Guidance for Industry



De Backer and Fay (to appear 2024)

3. Covariate adjustment

- Randomized trials: increase precision
- Observational trials: avoid bias

- Conditional estimate (equal covariates)

$$\mathrm{cPI} = \mathbb{P}\left(Y_i^T > Y_j^C \mid X^T = X^C\right) + \frac{1}{2}\mathbb{P}\left(Y_i^E = Y_j^C \mid X^T = X^C\right)$$

- Marginal estimate

$$\mathrm{mPI} = \mathbb{P}\left(Y_i^T > Y_j^C\right) + \frac{1}{2}\mathbb{P}\left(Y_i^E = Y_j^C\right)$$

Conditional Models

Single outcome: (G)PIM[1,2]: semi-parametric modelling framework

$$\text{logit}\left(\mathbb{P}\left(Y_i^T \geq_\tau Y_j^C \mid X^T = X^C\right)\right) = \beta_0 + \beta_{Trt} X_{Trt} + \beta_X'(X^T - X^C)$$

$$expit(\beta_{Trt}) = conditional\ PI$$

extended to multivariate outcomes in very specific cases[3] and for small sample and near-separation[4]

1. Thas et al. J R Stat Soc Series B Stat Methodol. (2012)
2. Zhang et al. International Statistical Review (2019)
3. Mao et al. Biometrics (2021)
4. Jaspers et al. Stat. Med. (2024)

Conditional Models: pim package

- Childhood Respiratory Disease Study (CRDS) follows the pulmonary function (FEV) in 654 children of ages 3–19.
- Interest: effect of smoking on FEV, corrected for age

```
> pim2 <- pim(FEV ~ Age*Smoke, data = FEVData)
> summary(pim2)
pim.summary of following model :
 FEV ~ Age * Smoke
Type:  difference
Link:  logit

           Estimate Std. Error z value Pr(>|z|)
Age         0.60760    0.03012  20.170  < 2e-16
Smoke       5.30689    1.04423   5.082 3.73e-07
Age:Smoke  -0.45539    0.07854  -5.798 6.71e-09
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.(
```

For 2 randomly selected children with the same smoking status and a year difference, the probability that the eldest has a higher FEV is estimated by:

$$\mathbb{P}\left(Y_i^T \geq_\tau Y_j^C | X_S^T = X_S^C, X_A^T = X_A^C + 1\right) = \frac{e^{0.61-0.46\,X_S}}{1+e^{0.61-0.46\,X_S}}$$

for $X_S = 0$:

= expit (0.61) = 0.65

for $X_S = 1$:

= expit (0.61-0.46) = 0.54

Rosner Fundamentals of Biostatistics. (1999)

Conditional Models: small sample pim

- Toxicology study:
  Investigate the impact of increasing doses of pyridine (0, 50,100, 250, 500, 1000 ppm) in 120 rats on organ weight corrected for sex and blood chemistry.
- 10 rats per gender & dose

Probability that weight gain is smaller in 0 ppm group compared to 500 ppm = 7.9%

		Male rats				
		50 ppm	100 ppm	250 ppm	500 ppm	1000 ppm
0 ppm	PI (%) (CI)	71.7 (36.68;91.72)	73.93 (27.7;95.45)	63.32 (26.21;89.35)	7.9 (1.5;32.56)	<0.01 (<0.01;<0.01)
	Unadj p	0.2142	0.3039	0.4946	0.0058	<0.0001
	Adj p	0.4015	0.5362	0.7809	0.0175	<0.0001
		Female rats				
		50 ppm	100 ppm	250 ppm	500 ppm	1000 ppm
0 ppm	PI (%)	42.78 (14.89;76.17)	60.83 (4.6;98.04)	45.23 (15.59;78.69)	9.18 (0.69;59.49)	1.04 (0.14;7.38)
	Unadj p	0.6921	0.8019	0.8005	0.0924	<0.0001
	Adj p	0.8591	0.8591	0.8591	0.2131	0.0001

Jaspers et al. Stat. Med. (2024)

Marginal Models

Single outcome:

- Regression imputation estimator[1]
- Inverse probability of treatment weighted (IPTW) estimator[2-4]

Working on extensions to multivariate outcomes

1. Vermeulen et al. Stat Med. (2015)
2. Vermeulen et al. Int J Biostat. (2016)
3. Mao et al. Biometrika (2018)
4. Zhang et al. International Statistical Review (2019)

Designing a Trial

Trial design for a benefit-risk question

Clinical situation

- Disease: low and intermediate risk acute promyelocytic leukemia (APL)
- Standard of care: high dose of ATRA (all-trans retinoic acid)
- Toxicity of ATRA remains a problem
- Real-world data suggest that a reduced dose may provide similar efficacy

→ **Non-inferiority trial?**

Control : Full ATRA dose
Experimental : Reduced ATRA dose

- Endpoint of interest: Event-Free Survival (EFS) at 2 years
- Sample size?

De Backer, Clin Trials 2023

Trial design for a benefit-risk question

Non-inferiority trial?

Outcome	Value in Control Arm*	Value in Experimental Arm**	Difference (Control – Experimental)	Non-inferiority margin	Approximate Sample Size
EFS at 2 years non-inferior	0.92	0.92	0.0	0.05	1000
	0.92	0.91	0.01	0.05	1500
	0.92	0.90	0.02	0.05	2900

De Backer, Clin Trials 2023

Trial design for a benefit-risk question

Use GPC as a pragmatic alternative

- Prioritized outcomes:
    1. EFS at 2 years of follow-up (alive and disease-free vs not)
    2. Grade 3/4 documented infections (no vs yes)
    3. Grade 3/4 differentiation syndrome (no vs yes)
    4. Grade 3/4 hepatotoxicity (no vs yes)
    5. Grade 3/4 neuropathy (no vs yes)
- Gains on toxicities are acceptable ***only if*** EFS results are similar.
- Efficacy $\succ$ toxicities

  $\Rightarrow$ EFS $\succ$ Tox1 $\succ \ldots \succ$ Tox4

De Backer, Clin Trials 2023

Trial design for a benefit-risk question

Key idea is to transform NI question into superiority question:

"Does benefit-risk balance favor reduced dose?"

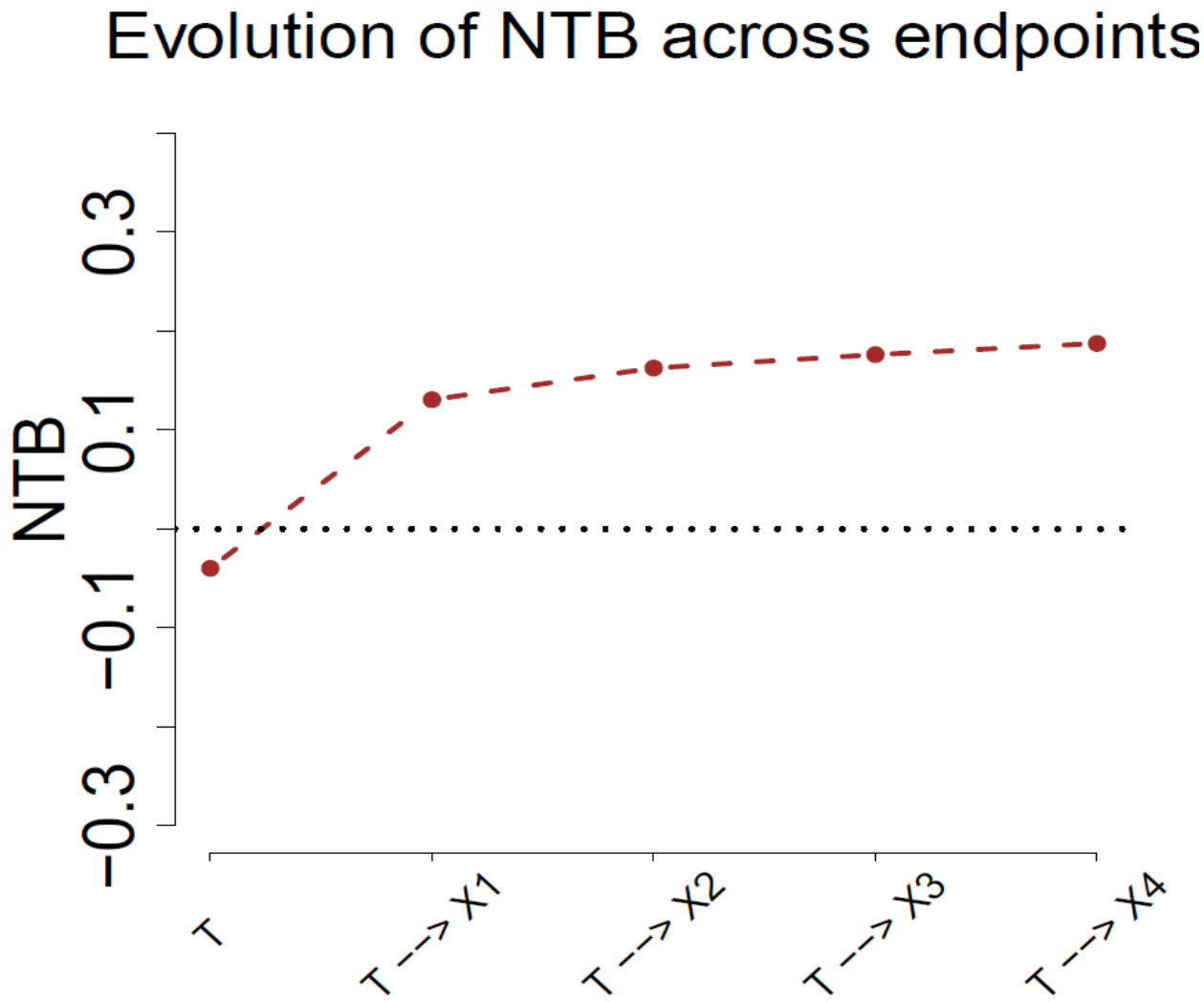
Use historical data to formulate assumptions about expected results for the outcomes with reduced dose.

Perform simulations to calculate NTB

Correlation between outcomes must be estimated from historical data

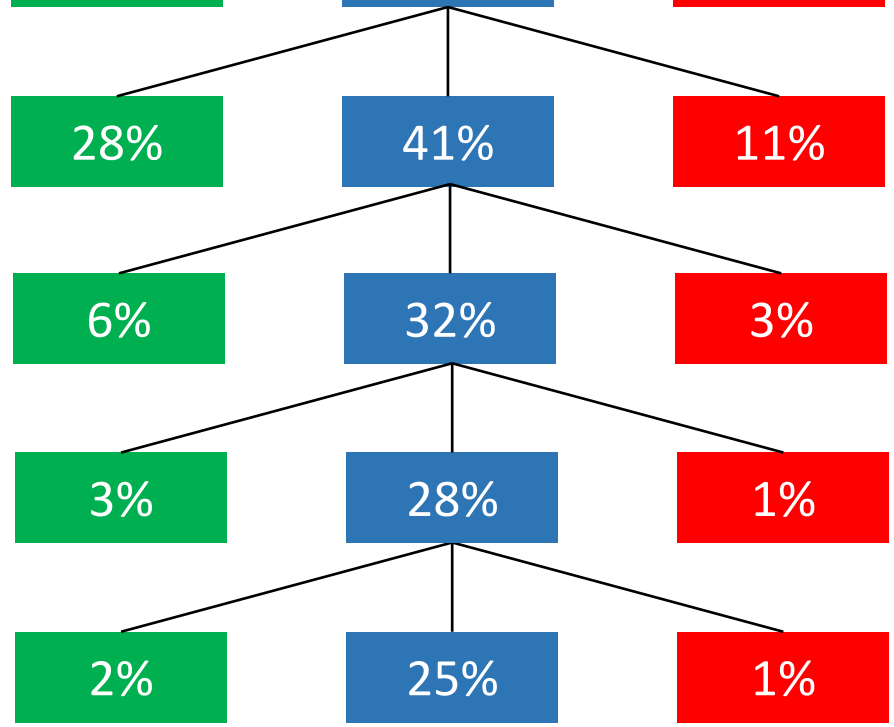
De Backer, Clin Trials 2023

Trial design for a benefit-risk question

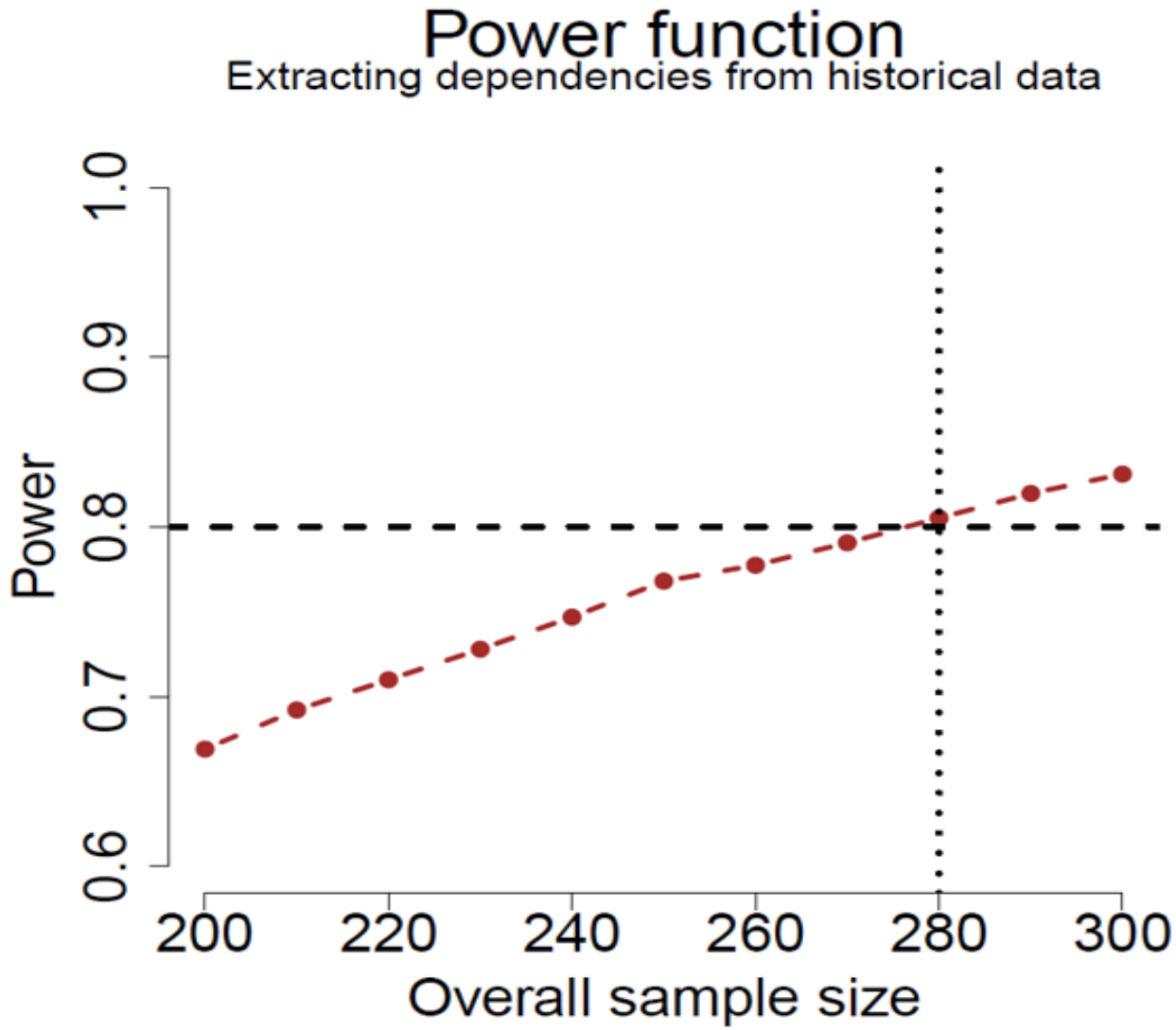


De Backer, Clin Trials 2023

Prioritized outcomes	Favorable pairs	Favor Low dose	Neutral	Favor High dose	Contrib	NTB
EFS @ 2 years	Higher	8%	80%	12%	- 4%	- 4%
Documented infection	Yes vs No	28%	41%	11%	17%	13%
Differentiation syndrome	Yes vs No	6%	32%	3%	3%	16%
Hepatotoxicity	Yes vs No	3%	28%	1%	2%	18%
Neuropathy	Yes vs No	2%	25%	1%	1%	19%
Overall		47%		28%		19%



Trial design for a benefit-risk question



De Backer, Clin Trials 2023

Trial design for a benefit-risk question

GPC accumulates evidence across all risks and benefits
→ smaller studies than to show non-inferiority on single risk

Power	**Traditional design (Non Inferiority)**	**GPC design (Net Benefit)**
80%	~750 patients	~310 patients
90%	~1010 patients	~410 patients

De Backer, Clin Trials 2023

U.S. FOOD & DRUG ADMINISTRATION

Benefit-Risk Assessment for New Drug and Biological Products

Guidance for Industry

U.S. FOOD & DRUG ADMINISTRATION

Patient-Focused Drug Development: Incorporating Clinical Outcome Assessments Into Endpoints For Regulatory Decision-Making

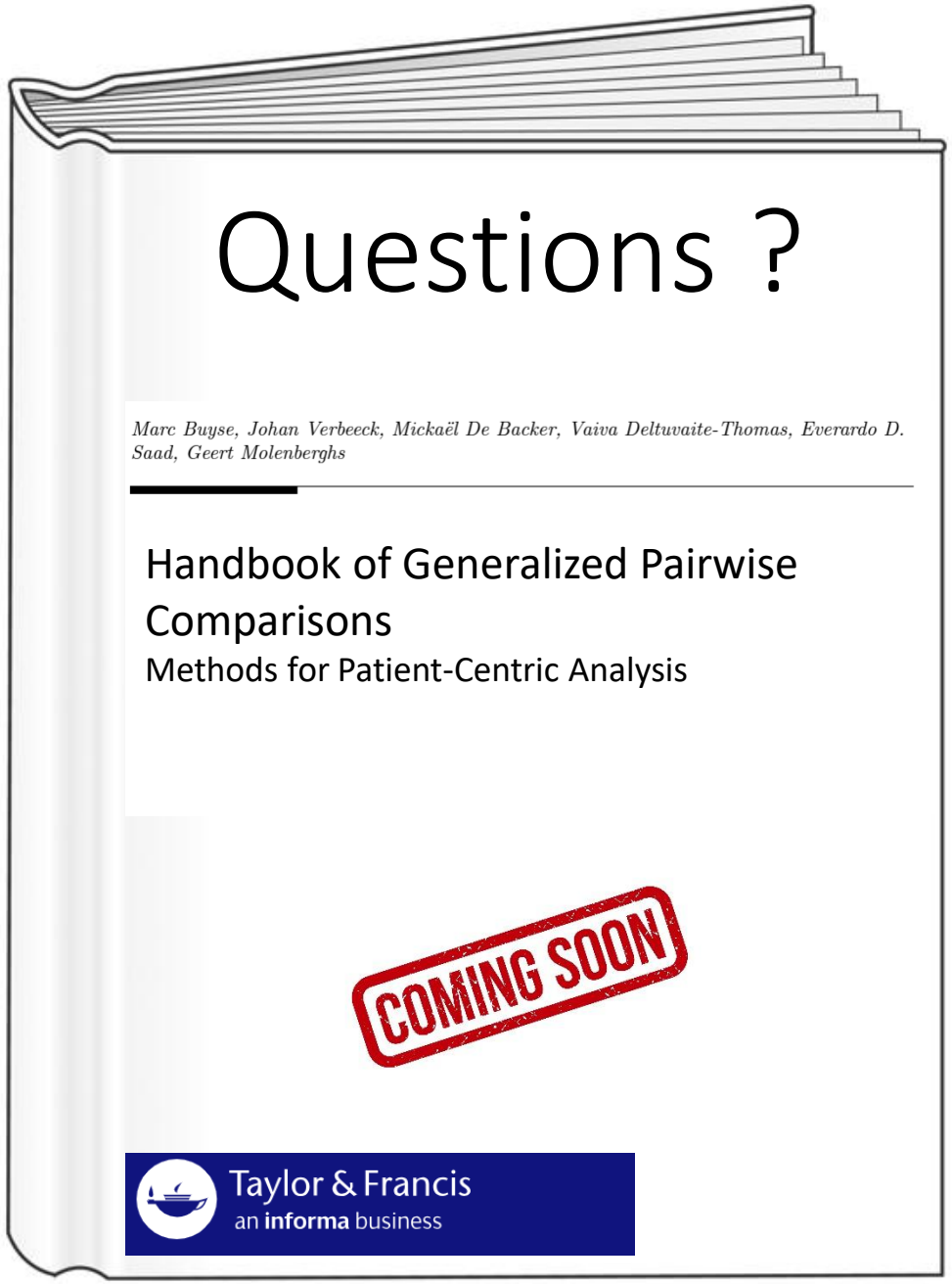
Guidance for Industry, Food and Drug Administration Staff, and Other Stakeholders

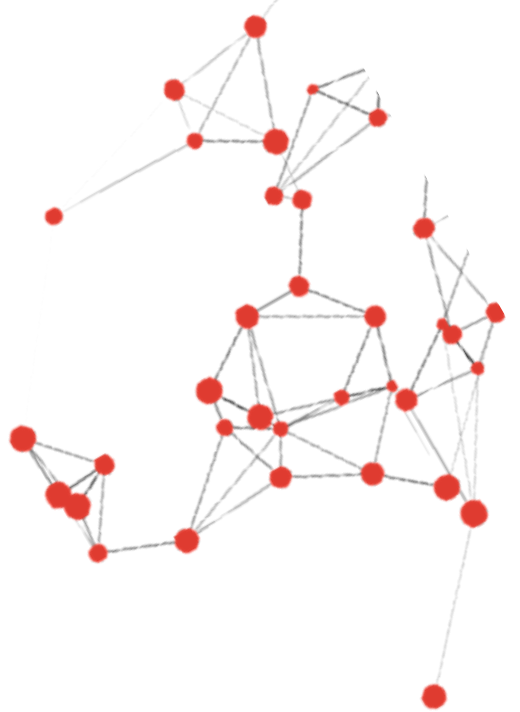
DRAFT GUIDANCE

GPC + / –

Advantages	Limitations
• Patient-centric	• Dangers of multiplicity
• Flexible	• Potential for data-driven analyses
• Multiple outcomes (*e.g.*, prioritized)	• More complex interpretation
• Higher power from multiple outcomes	• Lower power for benefit / risk analyses
• Thresholds of clinical relevance	• Lack of familiarity
• NTB easily interpretable effect measure	• WR not easily interpretable effect measure
• Absolute effects needed for benefit / risk	• Relative effects constant across subsets

Questions ?





Brice Ozenne PhD

brice.ozenne@nru.dk

Biostatistics & Neurobiology Research Unit

University of Copenhagen- Denmark

Johan Verbeeck PhD

johan.verbeeck@uhasselt.be

Data Science Institute

UHasselt - Belgium