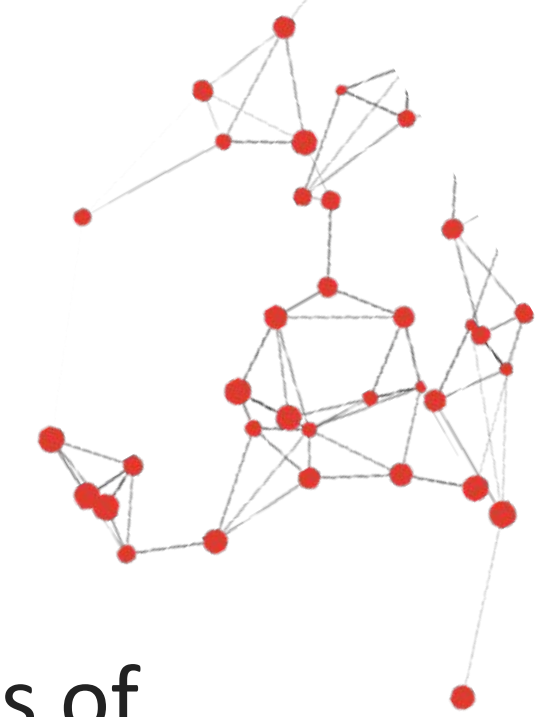


Generalized pairwise comparisons: A practical guide to the design and analysis of patient-centric trials

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. Benefit-risk assessment
  3. Multivariate outcomes of different types
- Concept of GPC

Agenda

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

- Introduction to the R package BuyseTest
  1. Measures of treatment effect
  2. Statistical inference
- Examples, revisited

Agenda

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

- Introduction to the  package BuyseTest
  1. Measures of treatment effect
  2. Statistical inference
- Examples, revisited

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

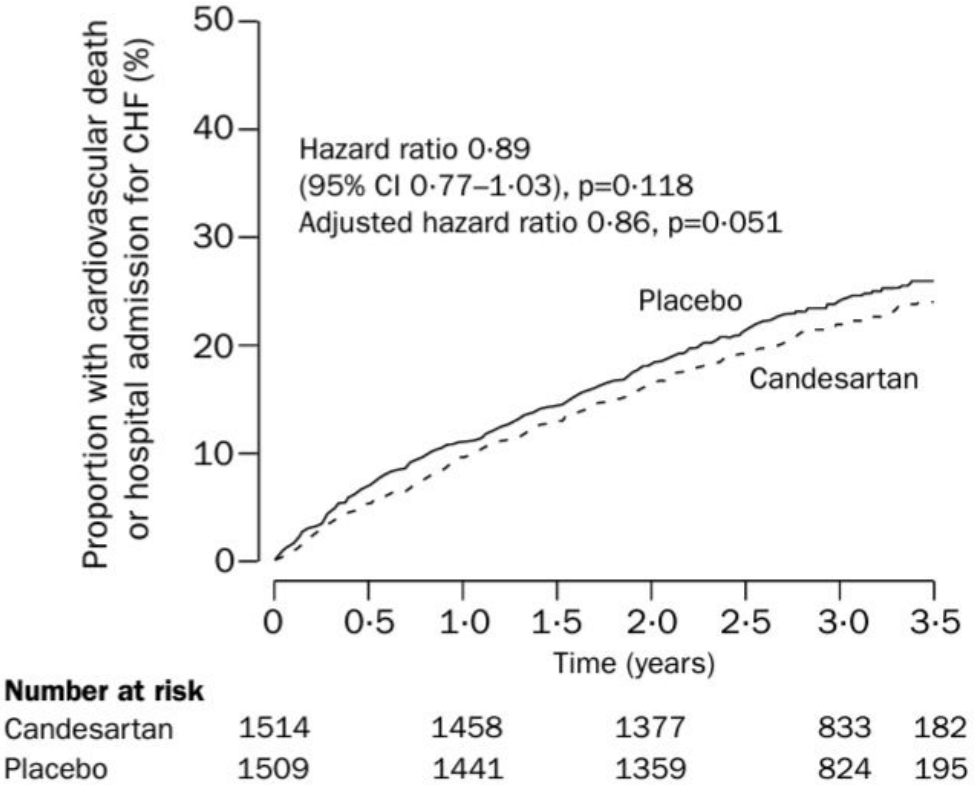
Motivating examples

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 event composite

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

46% (158/340) of CV deaths are ignored

Yusuf et al. Lancet (2003)

Issues with time-to-first event analyses

- **Emphasis is on time of event, rather than severity of event**; i.e. a patient that has an hospitalization is worse than a patient dying 1 day later

- **Ignores repeated events** (cannot count events; f.e. # hospitalizations)

	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. Benefit-Risk assessment

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Increased Survival in Pancreatic Cancer with nab-Paclitaxel plus Gemcitabine

Daniel D. Von Hoff, M.D., Thomas Ervin, M.D., Francis P. Arena, M.D.,

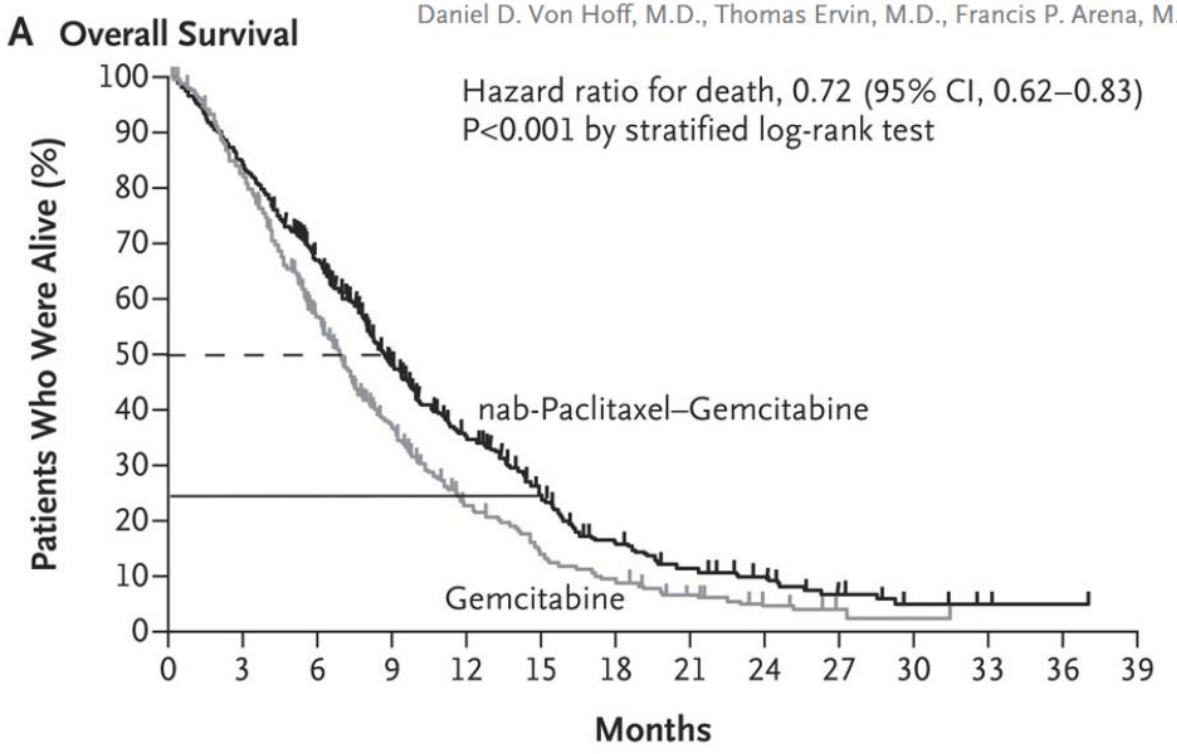


Table 3. Common Adverse Events of Grade 3 or Higher and Growth-Factor Use.*

Event	nab-Paclitaxel plus Gemcitabine (N=421)	Gemcitabine Alone (N=402)
Adverse event leading to death — no. (%)	18 (4)	18 (4)
Grade ≥3 hematologic adverse event — no./total no. (%)†		
Neutropenia	153/405 (38)	103/388 (27)
Leukopenia	124/405 (31)	63/388 (16)
Thrombocytopenia	52/405 (13)	36/388 (9)
Anemia	53/405 (13)	48/388 (12)
Receipt of growth factors — no./total no. (%)	110/431 (26)	63/431 (15)
Febrile neutropenia — no. (%)‡	14 (3)	6 (1)
Grade ≥3 nonhematologic adverse event occurring in >5% of patients — no. (%)‡		
Fatigue	70 (17)	27 (7)
Peripheral neuropathy§	70 (17)	3 (1)
Diarrhea	24 (6)	3 (1)
Grade ≥3 peripheral neuropathy		
Median time to onset — days	140	113
Median time to improvement by one grade — days	21	29
Median time to improvement to grade ≤1 — days	29	NR
Use of nab-paclitaxel resumed — no./total no. (%)	31/70 (44)	NA

- Do survival benefits outweigh the burden of the adverse-events?

Marginal Benefit-Risk analyses

Traditional analysis:

- Benefit: ~~log-rank test~~
  difference in 1 year survival

  example: 50% vs. 20%

- Risk: ~~chi-squared test~~
  difference in proportion of patients with events of grade 3 or higher

  example: 30% vs 0%

- ignores a possible association between benefit and risk
  - **a) positive association:** side effects may only occur when it prolongs life (never purely harmful treatment).
  - **b) no association:** treatment with two independent mechanisms, one acting on survival and another generating side effects.
  - **c) negative association:** treatment solely beneficial for some patients while solely harmful for other patients.

Benefit-Risk association

Treatment group (scenario a)		Response		
		Absent	Present	Total
Toxicity	Absent	0.5	0.2	0.7
	Present	0	0.3	0.3
	Total	0.5	0.5	1

benefit: 0.3 vs. 0.3

Control group		Response		
		Absent	Present	Total
Toxicity	Absent	0.8	0.2	1
	Present	0	0	0
	Total	0.8	0.2	1

- - - short life with toxicity	- short life without toxicity	+ long life with toxicity	+ + + long life without toxicity

Benefit-Risk association

Treatment group (scenario a)		Response		
		Absent	Present	Total
Toxicity	Absent	0.5	0.2	0.7
	Present	0	0.3	0.3
	Total	0.5	0.5	1

benefit: 0.3 vs. 0.3

Control group		Response		
		Absent	Present	Total
Toxicity	Absent	0.8	0.2	1
	Present	0	0	0
	Total	0.8	0.2	1

Treatment group (scenario b)		Response		
		Absent	Present	Total
Toxicity	Absent	0.35	0.35	0.7
	Present	0.15	0.15	0.3
	Total	0.5	0.5	1

unclear: 0.15 + 0.15 + 0.15 vs 0.45

Treatment group (scenario c)		Response		
		Absent	Present	Total
Toxicity	Absent	0.2	0.5	0.7
	Present	0.3	0	0.3
	Total	0.5	0.5	1

unclear: 0.3 + 0.3 vs 0.6

- - - short life with toxicity	- short life without toxicity	+ long life with toxicity	+ + + long life without toxicity

Sensible Benefit-Risk analyses

- Benefit-risk balance depends on the association between response and toxicity

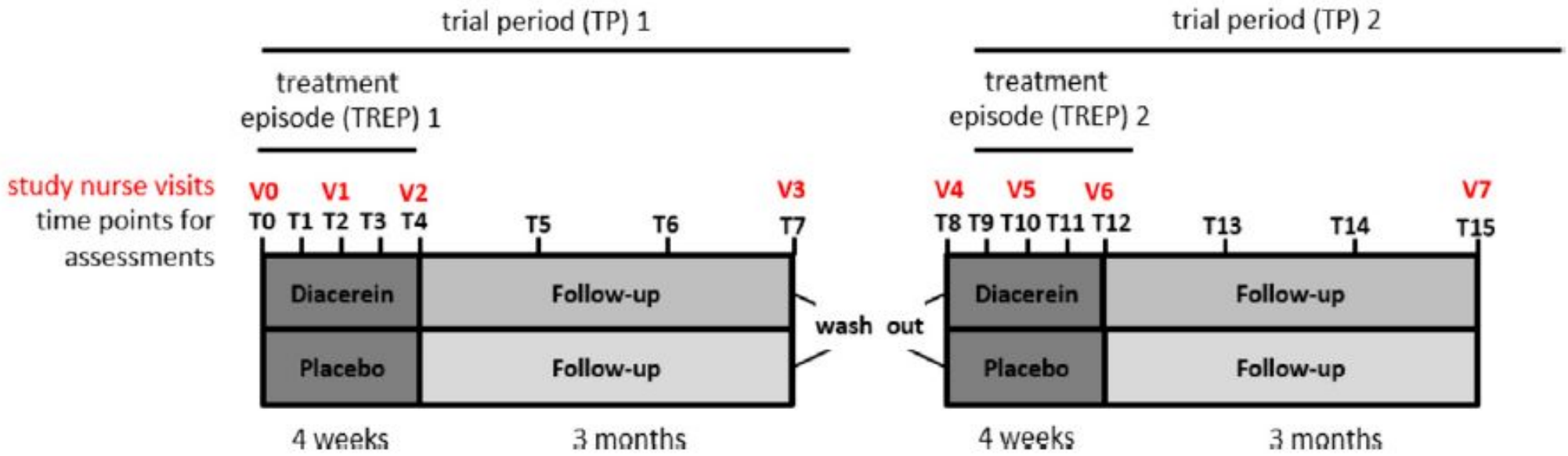
- Traditional (marginal) analysis ignore this association
  - marginal benefits and marginal risks cannot be combined (without strong assumptions)
  - interpretation of the results is difficult

- Upon deciding on a hierarchy of outcomes, e.g.:

a joint analysis of the benefit and the risk will provide value information.

3. Multivariate outcomes of different types

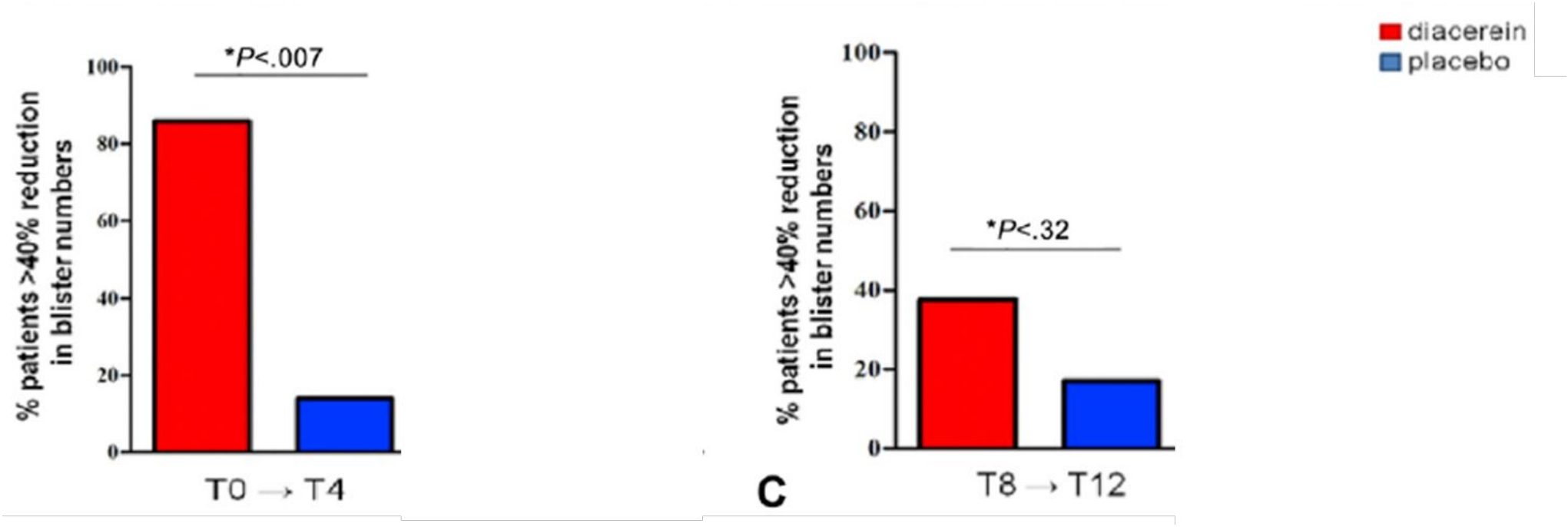
- Rare skin disease: Epidermolysis bullosa simplex
- Formation of blisters under low mechanical stress
- 16 pediatric subjects treated with placebo and diacerin cream in a longitudinal cross-over trial



Wally et al. J Am Acad Dermato (2018)

Inconclusive results - primary endpoint analysis

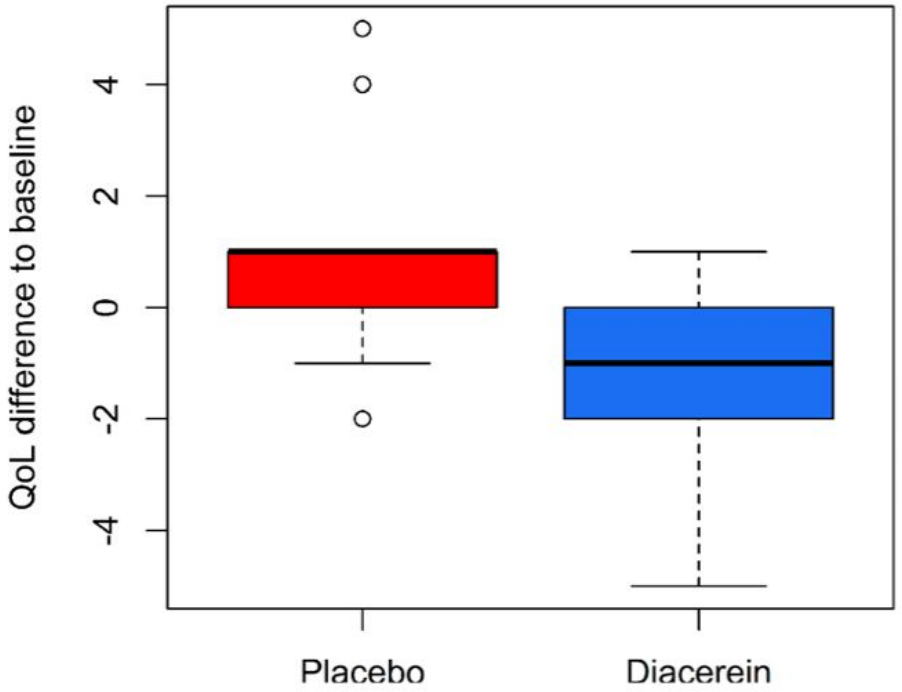
- Primary endpoint:
  - >40% reduction in blisters compared to baseline (binary outcome) at week 4
  - Barnard test (~Fisher exact test 2x2 table) per treatment period

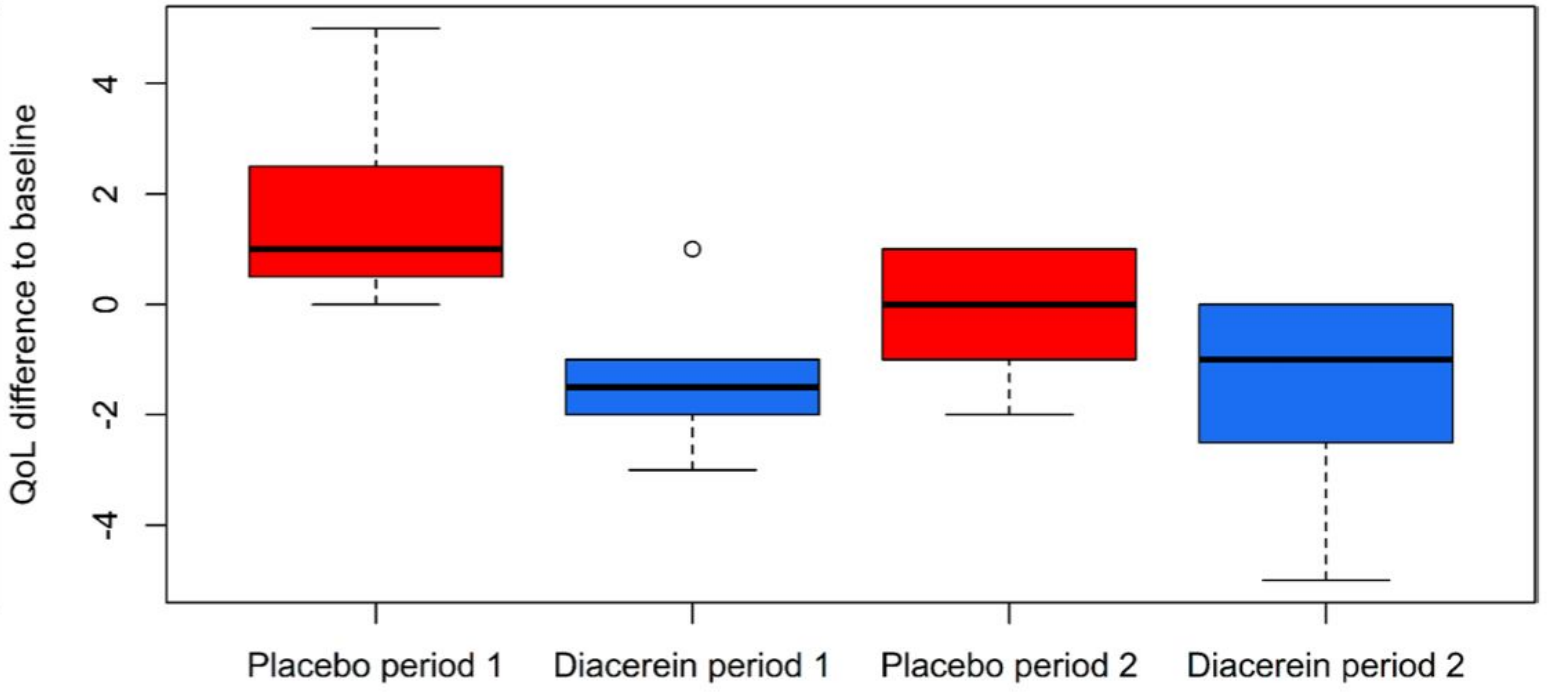


Wally et al. J Am Acad Dermato (2018)

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)

Concept of GPC

Wilcoxon rank-sum test

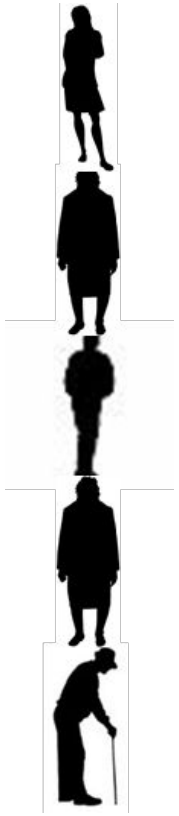
New Treatment

Group E

$i = 1, \ldots, n^E$

Comparator

Group C



$j = 1, \ldots, n^C$

1. Order the $Y^E \cup Y^C$ elements ($\mathrm{n} = n^E + n^C$)

2. Let R_n be the rank order of the n^{th} element

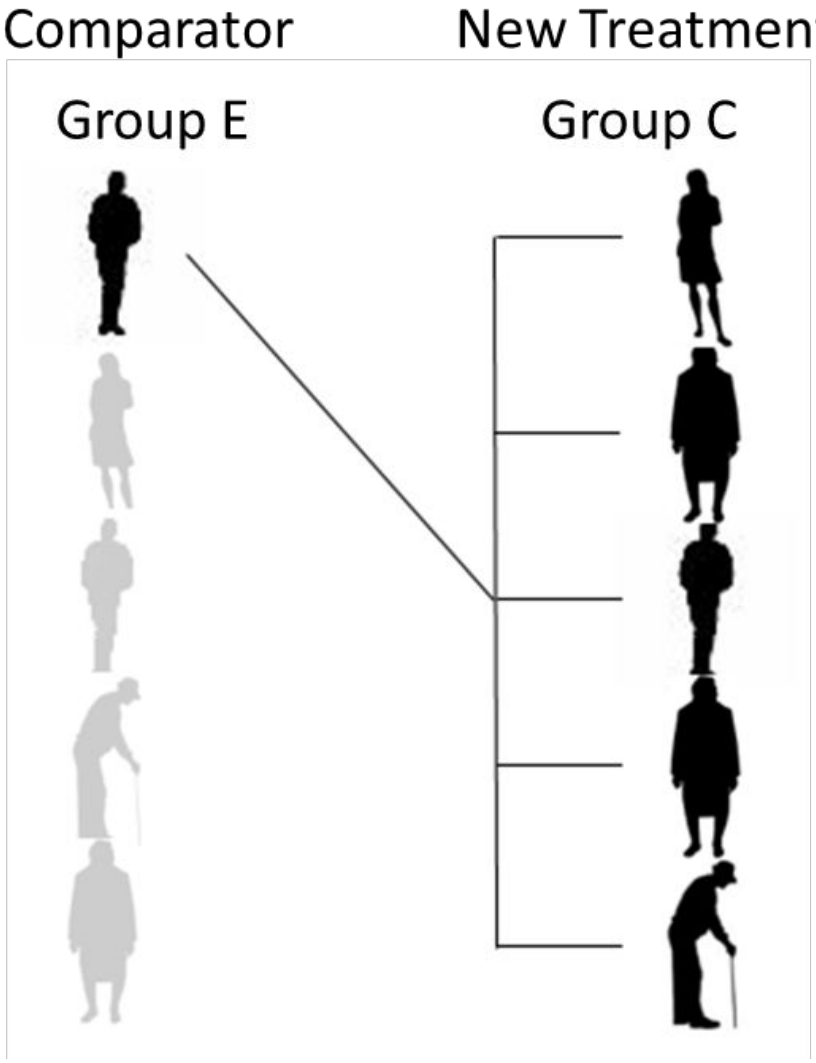
3. Calculate the sum of the ranks of Y^E:

$$\widehat{U} = \sum_{i=1}^{n^E} R_i$$

4. The statistic $\widehat{U}$ has a known distribution under H_0

Wilcoxon. Biometrics (1945)

Mann-Whitney test



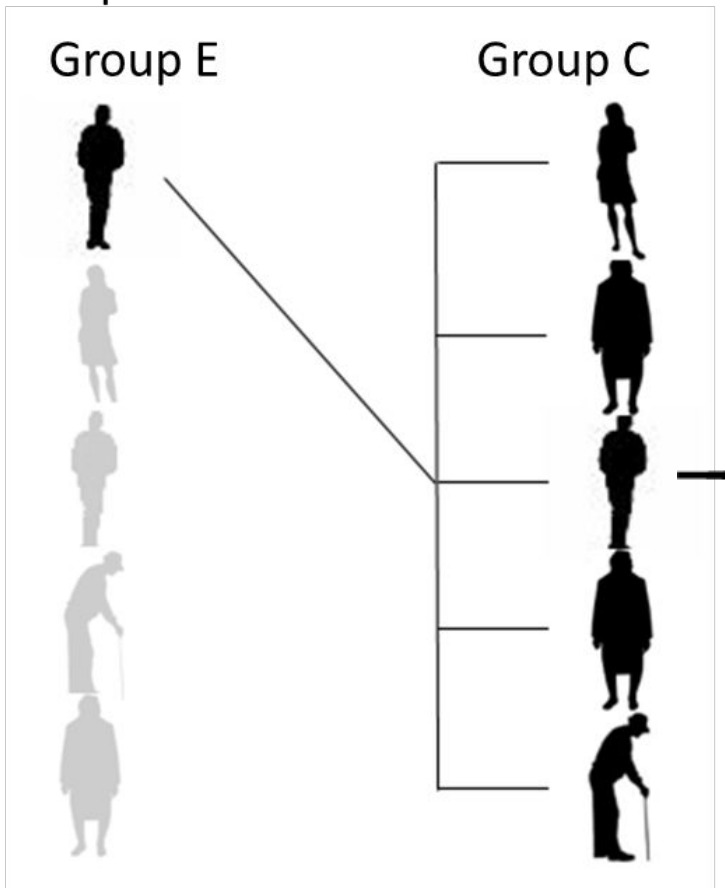
1. Perform pairwise comparisons between all elements of Y^E and Y^C

2. Calculate $U_{ij}^{MW} = \begin{cases} 1 & if\ Y_i^E > Y_j^C \\ 0 & if\ Y_i^E < Y_j^C \\ 1/2 & if\ Y_i^E = Y_j^C \end{cases}$

3. The statistic $\widehat{\Delta}^{MW} = \frac{1}{n^E n^C} \sum_{i=1}^{n^E} \sum_{j=1}^{n^C} U_{ij}^{MW}$ has a known distribution under H_0

Mann & Whitney. Ann Math Stat (1947)

Generalized Pairwise Comparisons (GPC)



Highest ranked outcome

tie

Middle ranked outcome

tie

Lowest ranked outcome

1. Perform pairwise comparisons between all elements of Y^E and Y^C

2. Calculate $U_{ij} = \begin{cases} 1 & if\ Y_i^E \succ Y_j^C \\ -1 & if\ Y_i^E \prec Y_j^C \\ 0 & if\ Y_i^E \asymp Y_j^C \end{cases}$

3. The statistic $\widehat{\Delta} = \frac{1}{n^E n^C} \sum_{i=1}^{n^E} \sum_{j=1}^{n^C} U_{ij}$ has a known distribution under H_0

Note: priorities may be patient-centric

Buyse. Stat Med (2010)
Pocock et al. Eur Heart J (2012)

GPC – threshold of clinical similarity

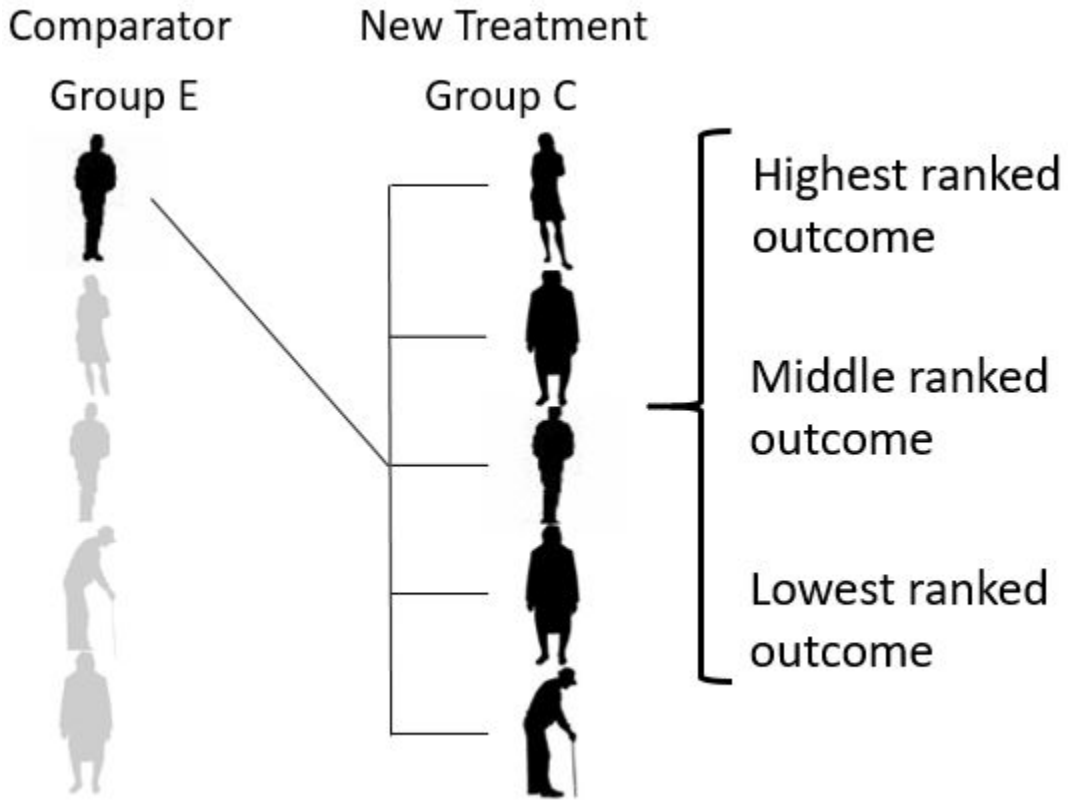
1. Perform pairwise comparisons between all elements of Y^E and Y^C

2. Calculate $U_{ij} = \begin{cases} 1 & if\ Y_i^E > Y_j^C + \tau \\ -1 & if\ Y_i^E < Y_j^C + \tau \\ 0 & otherwise \end{cases}$

3. The statistic $\widehat{\Delta}_\tau = \frac{1}{n^E n^C} \sum_{i=1}^{n^E} \sum_{j=1}^{n^C} U_{ij}$ has a known distribution under H_0

Buyse. Stat Med (2010)

GPC – multiple weighted outcomes



1. Perform pairwise comparisons between all elements of Y^E and Y^C for each of the k =(1,..,d) outcomes

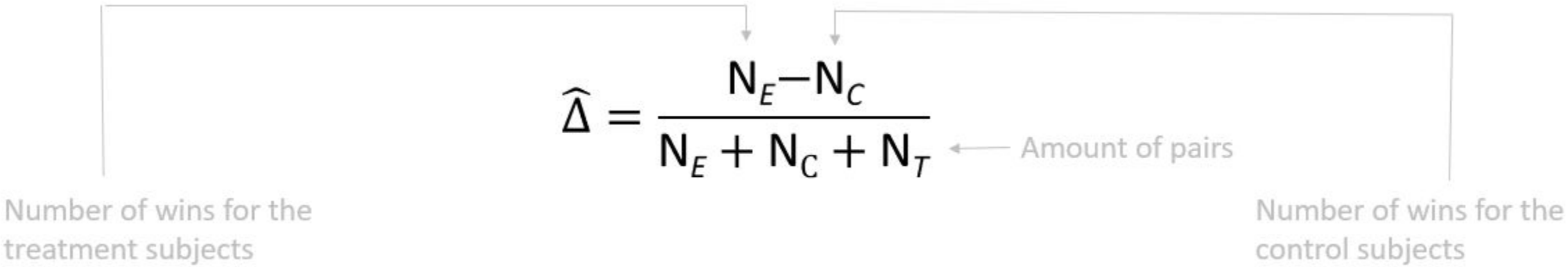
2. Calculate $U_{ij}(k) = \begin{cases} 1 & if\ Y_i^E > Y_j^C \\ -1 & if\ Y_i^E < Y_j^C \\ 0 & if\ Y_i^E = Y_j^C \end{cases}$

3. The statistic $\widehat{\Delta}^O = \frac{1}{n^E n^C} \sum_{i=1}^{n^E} \sum_{j=1}^{n^C} \sum_{k=1}^{d} w(k)\, U_{ij}(k)$ has a known distribution under H_0

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

O'Brien. Biometrics (1984)

GPC statistics



Net treatment benefit (NTB)

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

Is a U-statistic

Buyse. Stat Med (2010)
Hoeffding. Ann Math Stat (1948)

GPC statistics - Net treatment benefit

Estimates net treatment benefit (NTB): $\Delta = P(Y_E > Y_C) - P(Y_E < Y_C)$

Related to probabilistic index, Mann-Whitney effect,… (θ):

$$\theta = P(Y_E > Y_C) + \frac{1}{2}P(Y_E = Y_C)$$

$$\Delta = 2\,\theta - 1$$

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.

Buyse. Stat Med (2010)

GPC statistics

$$\widehat{\Delta} = \frac{N_E - N_C}{N_E + N_C + N_T}$$

Amount of pairs

Number of wins for the treatment subjects

Number of wins for the control subjects

$$\text{Success Odds} = \frac{N_E + {}^1/_2 N_T}{N_C + {}^1/_2 N_T}$$

Number of ties

$$\text{Win Ratio} = \frac{N_E}{N_C}$$

Note : NTB = (SO-1)/(SO+1)

Buyse. Stat Med (2010)
Pocock et al. Eur Heart J (2012)
Dong et al. Stat Biopharm Res (2019)
Brunner et al. Stat Med (2021)

GPC statistics - differences

	Wins N_E (%)	Losses N_C (%)	Ties N_T (%)	NTB $\frac{N_E - N_C}{N_E + N_C + N_T}$	SO $\frac{N_E + 0.5N_T}{N_C + 0.5N_T}$	WR $\frac{N_E}{N_C}$
Trial 1	3 (0.06%)	1 (0.02%)	4,996 (99.92%)	0.0004	1.0008	3.00
Trial 2	3,000 (60%)	1,000 (20%)	1,000 (20%)	0.40	2.33	3.00

The WR ignores the ties or redistributes the ties according to the observed win/loss proportions -> overestimation of effect

Consequences of redistributing ties

Continuous

- f.e. relative change in NT-proBNP (PARACHUTE-HF)
- Chance of a tie is negligible; unless you use a threshold

Discrete

- f.e. count (# of hospitalisations), categorical (Yes/No, QoL, 6MWT improvement,..)
- Chance of a tie is very high
- WR redistributes ties according to win proportions of not-tied pairs

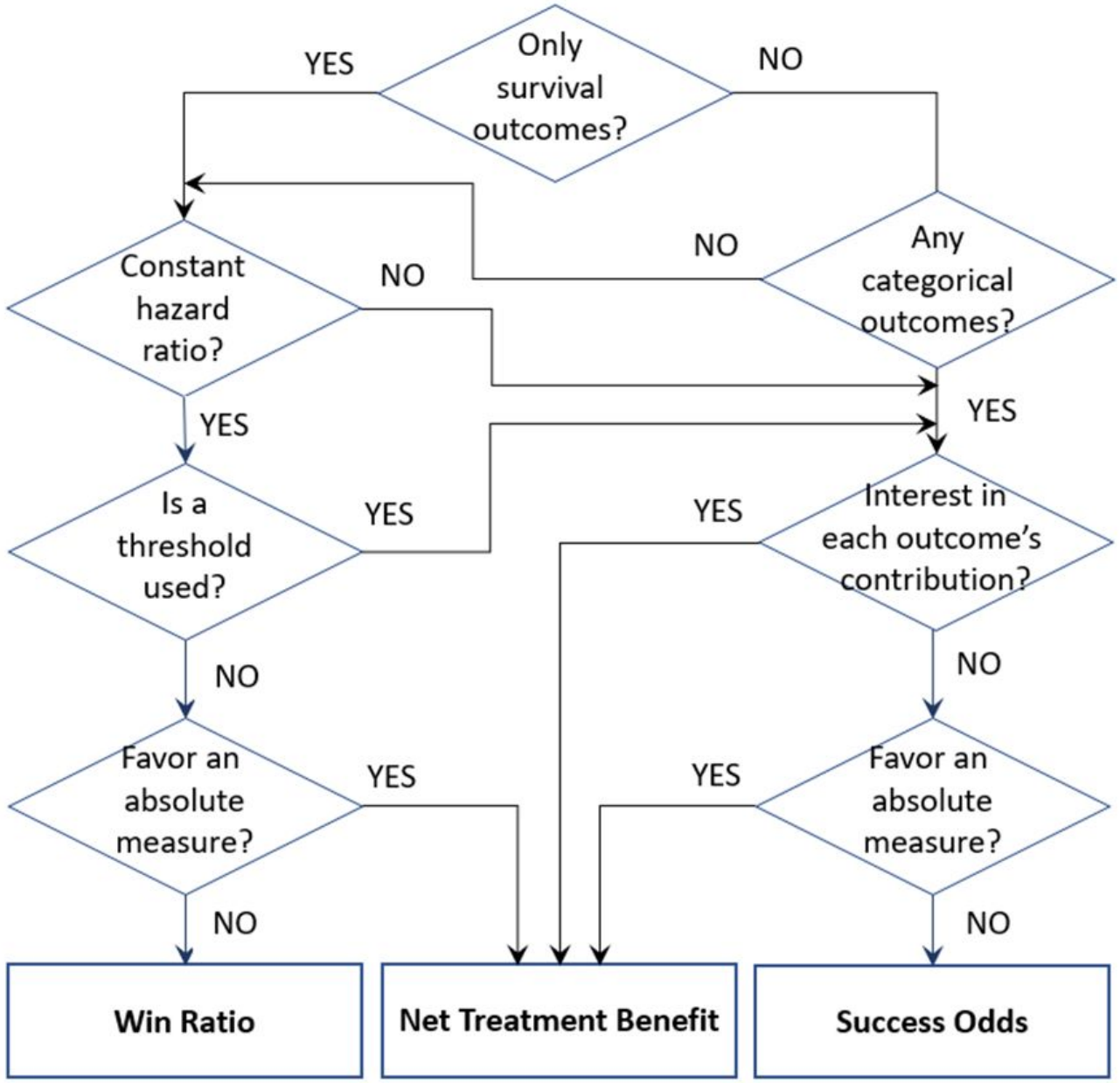
Survival

- f.e. equal time to death or censored death time
- Chance of a tie due to censored event time is very common; due to equal event time uncommon
- WR assumes that censored events will behave as observed events = ok under proportional hazards, but unrealistic on both joint and individual survival outcome

Additive decomposition of NTB

365 days	% wins	% losses	% ties	NTB (95%CI)	SO (95%CI)	WR (95%CI)
Death	4.31	3.62	92.07	0.0069	1.01	1.19
Hemor. Stroke	0.05	0.07	91.95	-0.0002	1.00	0.67
Isch. Stroke	0.41	0.29	91.25	0.0011	1.00	1.39
MI	9.70	8.90	72.65	0.0080	1.02	1.09
Total MACE	14.47	12.88	72.65	0.016 (0.000-0.031)	1.03 (1.00-1.06)	1.12 (1.00-1.26)
p-value MACE				0.0413	0.0413	0.0414

Guidance of GPC measures in clinical trials



Verbeeck et al. JACC (2023)

GPC – Links with conventional effect size measures for univariate outcomes

Binary endpoint (denote success by 1 and failure by 0)

$$\Delta = P_E - P_C \qquad \text{WR} = \frac{P_E/(1-P_E)}{P_C/(1-P_C)}$$

Continuous endpoint

$$\Delta = 2\Phi\left(\frac{d}{\sqrt{2}}\right) - 1, \text{ with } d = \text{Cohen's d}$$

Survival endpoint (denote $\delta^E = \delta^C = 0$ as censored and $\delta^E = \delta^C = 1$ as observed event)

$$U_{ij} = \begin{cases} 1, & \text{if } Y_i^E > Y_j^C, \text{ and } \delta_j^C = 1 \\ -1, & \text{if } Y_i^E < Y_j^C, \text{ and } \delta_i^E = 1 \\ 0, & \text{if } Y_i^E = Y_j^C, \text{ and } \delta_i^E = \delta_j^C = 1 \\ 0, & \text{otherwise.} \end{cases} \qquad \Delta = \frac{1-HR}{1+HR} \qquad \text{WR} = \frac{1}{HR}$$

Gehan. Biometrika (1965)

Questions?

Break

Software for GPC

Part 1: Measures of treatment effect

code available at https://github.com/bozenne/tutorial-DagStat2025-GPC
file BuyseTest_intro.R

-package: BuyseTest

Installation
(requires internet connection)

R code

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

R output

```
Loading required package: Rcpp
BuyseTest version 3.1.0
```

Simulate data

- 100 subjects
- 2 treatment groups (C or T)
- outcomes simulated independently
- data.table format

(`as.data.frame`: convertion to data.frame)

R code

```
> set.seed(10)
> data <- simBuyseTest(100, n.strata = 2)
> head(data)
```

R output

```
      id treatment  eventtime status toxicity      score strata
   <num>    <fctr>      <num>  <num>   <fctr>      <num> <fctr>
1:     1         C 0.17392093      1      yes -2.1250686      a
2:     2         C 0.16255166      0      yes  0.5211787      a
3:     3         C 0.08302502      1      yes -0.0464229      b
4:     4         C 0.22204972      0       no -1.1494717      b
5:     5         C 0.11669726      1       no  0.6293383      a
6:     6         C 0.11885540      1      yes -0.7264715      a
```

Formula interface

Left hand side: group

Right hand side: outcome(s)

- `time to event` or `tte` or `t`
- `continuous` or `cont` or `c`
- `binary` or `bin` or `b`

R code

```
> e.BT <- BuyseTest(treatment ~ tte(eventtime, status = status),
                    data = data)
```

R output

```
        Generalized Pairwise Comparisons

Settings
- 2 groups  : Control = C and Treatment = T
- 1 endpoint:
priority endpoint   type           operator             event
1        eventtime  time to event  higher is favorable  status (0 1)
- right-censored pairs: probabilistic score based on the survival curves

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
```

Output of the GPC procedure

Default: net treatment benefit

$$\mathbb{P}\left[Y^E > Y^C\right] - \mathbb{P}\left[Y^C > Y^E\right]$$

R code

```
> summary(e.BT)
```

R output

```
        Generalized pairwise comparisons with 1 endpoint

- statistic       : net treatment benefit  (delta: endpoint specific, Delta: global)
- null hypothesis : Delta == 0
- confidence level: 0.95
- inference       : H-projection of order 1 after atanh transformation
- treatment groups: T (treatment) vs. C (control)
- censored pairs  : probabilistic score based on the survival curves
- results
endpoint total(%) favorable(%) unfavorable(%) neutral(%) uninf(%)  Delta CI [2.5% ; 97.5%] p.value
eventtime     100        57.39          42.61          0        0 0.1479  [-0.0293;0.3161] 0.10151
```

Display number of pairs instead of %:

R code

```
> summary(e.BT, percentage = FALSE)
```

Other measures of treatment effect

Win ratio

- argument statistic

R code

```
> confint(e.BT, statistic = "winRatio")
```

$$\frac{\mathbb{P}\left[Y^E > Y^C\right]}{\mathbb{P}\left[Y^C > Y^E\right]}$$

R output

```
          estimate        se  lower.ci upper.ci null   p.value
eventtime 1.347081 0.2450411 0.9430953 1.924118    1 0.1014458
```

Probabilistic index & success odds

$$\mathbb{P}\left[Y^E > Y^C\right] + 0.5\mathbb{P}\left[Y^E = Y^C\right]$$

re-run GPC adding the contribution of neutral pairs

R code

```
> e.BThalf <- BuyseTest(treatment ~ tte(eventtime, status),
                        data = data, add.halfNeutral = TRUE, trace = FALSE)
> model.tables(e.BThalf, statistic = "favorable")
```

R output

```
   endpoint total favorable unfavorable neutral uninf     Delta  lower.ci  upper.ci   p.value
1 eventtime   100  57.39388    42.60612       0     0 0.5739388 0.4852354 0.6581263 0.1019135
```

no neutral pairs here

so win ratio = success odds

R code

```
> coef(e.BThalf, statistic = "winRatio")
```

$$\frac{\mathbb{P}\left[Y^E > Y^C\right] + 0.5\mathbb{P}\left[Y^E = Y^C\right]}{\mathbb{P}\left[Y^C > Y^E\right] + 0.5\mathbb{P}\left[Y^C = Y^E\right]}$$

R output

```
[1] 1.347081
```

More options

Multiple outcomes

- separated by “+”
- priority from left (highest) to right (lowest)

Threshold (τ)

- time to event outcomes
- continuous outcomes

Operator

- “<0” lower values are favorable

R code

```
> e.MBT <- BuyseTest(treatment ~ tte(eventtime, status, threshold = 1) + bin(toxicity, operator = "<0"),
+                    data = data, trace = 0)
> model.tables(e.MBT)
```

$$\Delta_1 = \delta_1$$
$$\Delta_2 = \delta_1 + \delta_2$$

R output

	endpoint	threshold	total	favorable	unfavorable	neutral	uninf	delta	Delta	lower.ci	upper.ci	p.value
1	eventtime	1e+00	100.0	10.2	2.55	87.2	0	0.0768	0.0768	-0.00928	0.162	0.0803
3	toxicity	1e-12	87.2	18.8	24.72	43.7	0	-0.0590	0.0178	-0.13396	0.169	0.8192

$$\delta_1 = \mathbb{P}\left[Y_1^E > Y_1^C + \tau_1\right] - \mathbb{P}\left[Y_1^C > Y_1^E + \tau_1\right]$$
$$\delta_2 = \mathbb{P}\left[Y_2^E > Y_2^C \middle| |Y_1^E - Y_1^C| \leq \tau_1\right] - \mathbb{P}\left[Y_2^C > Y_2^E \middle| |Y_1^E - Y_1^C| \leq \tau_1\right]$$

Software for GPC

Part 2: statistical inference


code available at https://github.com/bozenne/tutorial-DagStat2025-GPC
file BuyseTest_intro.R

Inferential methods in Buyse Test

$H_0: F_e = F_c$
outcome distribution

permutation

$H_0: \Delta = 0$
Net Treatment Benefit

studentized permutation

no CI but exact
(under exchangeability)

u-statistic

bootstrap

u-statistic
+ transformation

studentized bootstrap

asymptotically valid

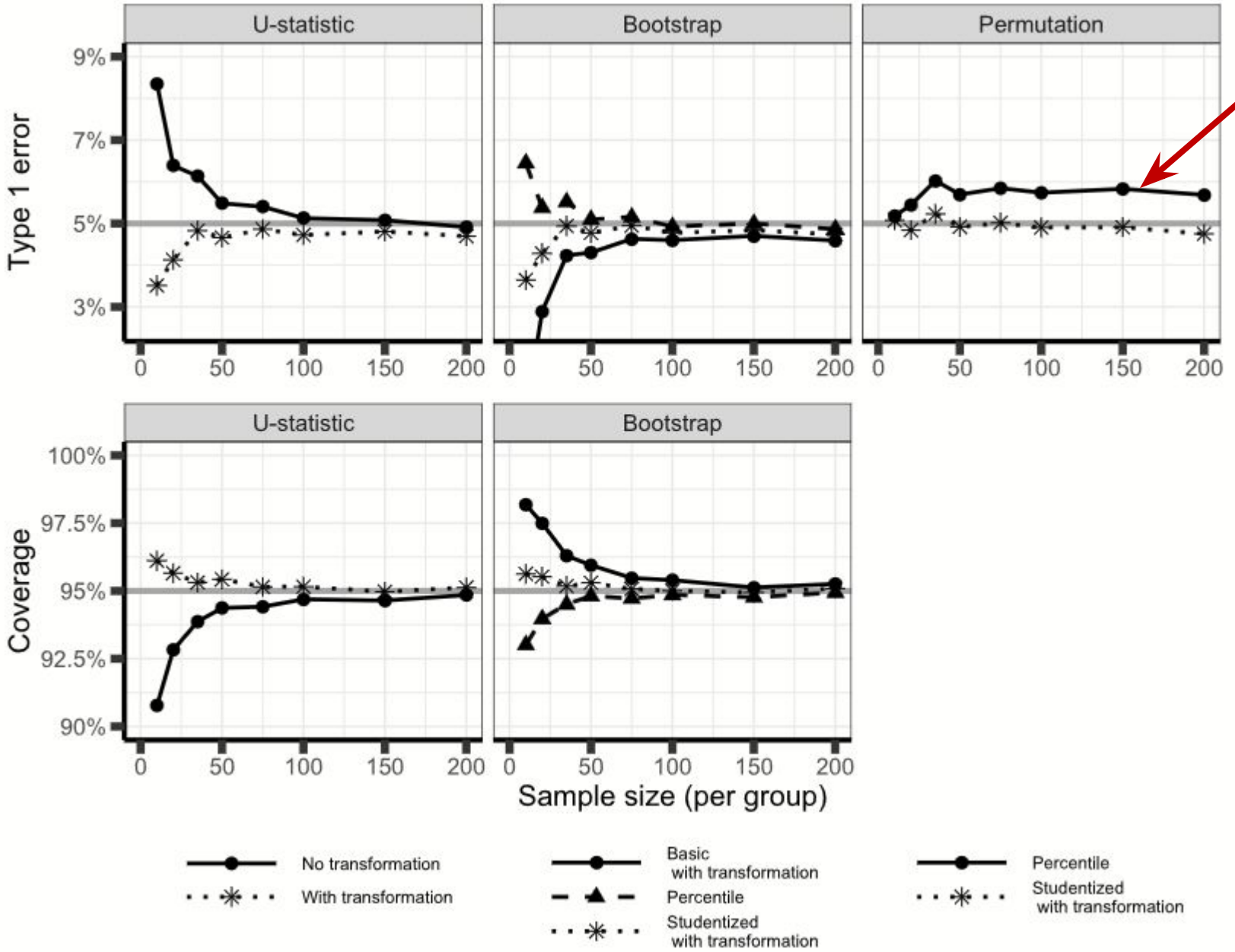
studentized: Δ/σ_Δ

Anderson & Verbeeck
Mathematics 2023

resampling (slow)

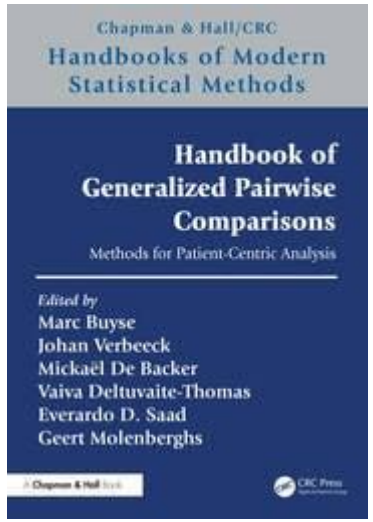
analytic (fast)

Results from a simulation study



heteroscedastic outcome

From



chapter 3

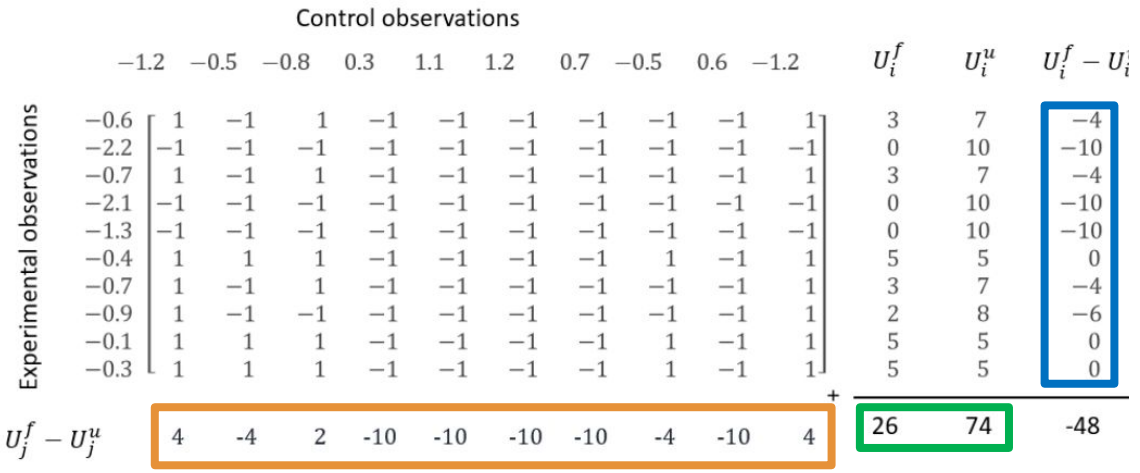
Inference based on U-statistic theory

Toy example

- 10 observation in each arm

Net Treatment Benefit (Δ)

- estimated U=(26-74)/100=-0.48



Theory: H-decomposition

- first order: similar to jackknife
- second order: asymptotically neglectable

$$U - \Delta = \underbrace{\frac{1}{m}\sum_{i=1}^{m} h_E(i)}_{\text{Experimental group}} + \underbrace{\frac{1}{n}\sum_{j=1}^{n} h_C(j)}_{\text{Control group}} + \underbrace{\frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} h_{EC}(i,j)}_{\text{Second order term}}$$

$$\text{where for } i \in \{1,\dots,m\}, h_E(i) = \mathbb{E}[\mathbb{1}_{Y_i^E > Y_j^C} - \mathbb{1}_{Y_j^C > Y_i^E} \mid Y_i^E] - \Delta$$

$$j \in \{1,\dots,n\}, h_C(j) = \mathbb{E}[\mathbb{1}_{Y_i^E > Y_j^C} - \mathbb{1}_{Y_i^C > Y_i^E} \mid Y_j^C] - \Delta$$

Ozenne et al. SMMR (2021)

Inference based on U-statistic theory

Variance estimation:

- first order:
  1.536/100+3.376/100=0.04912
- second order:
  9*first order/10 + … = 0.0519
  ⚠ second order term in presence of ties
  see Brunner et al. Stat. Papers (2025).

$$U - \Delta = \underbrace{\frac{1}{m}\sum_{i=1}^{m} h_E(i)}_{\text{Experimental group}} + \underbrace{\frac{1}{n}\sum_{j=1}^{n} h_C(j)}_{\text{Control group}} + \underbrace{\frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} h_{EC}(i,j)}_{\text{Second order term}}$$

$$\text{where for } i \in \{1,\dots,m\}, h_E(i) = \mathbb{E}[\mathbb{1}_{Y_i^E > Y_j^C} - \mathbb{1}_{Y_j^C > Y_i^E} \mid Y_i^E] - \Delta$$

$$j \in \{1,\dots,n\}, h_C(j) = \mathbb{E}[\mathbb{1}_{Y_i^E > Y_j^C} - \mathbb{1}_{Y_i^C > Y_i^E} \mid Y_j^C] - \Delta$$

$$\widehat{\sigma}_U \underbrace{\approx}_{\text{First order}} \frac{1}{m^2}\sum_{i=1}^{m} h_E^2(i) + \frac{1}{n^2}\sum_{j=1}^{n} h_C^2(j)$$

Implementation:

- combine intermediate results (nearly no extra calculations)
- use $\tanh^{-1}$ transformation to be range preserving (better small sample performance)

Control observations

Experimental observations	−1.2	−0.5	−0.8	0.3	1.1	1.2	0.7	−0.5	0.6	−1.2	U_i^f	U_i^u	$U_i^f - U_i^u$	$\left(\frac{U_i^f - U_i^u}{n^C} - \hat{U}\right)^2$
−0.6	1	−1	1	−1	−1	−1	−1	−1	−1	1	3	7	−4	0.0064
−2.2	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	0	10	−10	0.2704
−0.7	1	−1	1	−1	−1	−1	−1	−1	−1	1	3	7	−4	0.0064
−2.1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	0	10	−10	0.2704
−1.3	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	0	10	−10	0.2704
−0.4	1	1	1	−1	−1	−1	−1	1	−1	1	5	5	0	0.2304
−0.7	1	−1	1	−1	−1	−1	−1	−1	−1	1	3	7	−4	0.0064
−0.9	1	−1	−1	−1	−1	−1	−1	−1	−1	1	2	8	−6	0.0144
−0.1	1	1	1	−1	−1	−1	−1	1	−1	1	5	5	0	0.2304
−0.3	1	1	1	−1	−1	−1	−1	1	−1	1	5	5	0	0.2304
										+				
$U_j^f - U_j^u$	4	-4	2	-10	-10	-10	-10	-4	-10	4	26	74	-48	1.536
$\left(\frac{U_j^f - U_j^u}{n^E} - \hat{U}\right)^2$	0.7744	0.0064	0.4624	0.2704	0.2704	0.2704	0.2704	0.0064	0.2704	0.7744	+ 3.376			

Inference based on U-statistic theory (code)

R code

```
> rbind(confint(e.BT, transformation = TRUE),
        confint(e.BT, transformation = FALSE))
```

R output

```
           estimate         se    lower.ci  upper.ci null    p.value
eventtime  0.1478776 0.08897931 -0.02931684 0.3160612    0 0.10150573
eventtime1 0.1478776 0.08897931 -0.02651861 0.3222739    0 0.09652625
```

R code

```
> NTB <- coef(e.BT)
> sigma.NTB <- sqrt(crossprod(getIid(e.BT)))
> sigmaTrans.NTB <- sigma.NTB/(1-NTB^2)
> c(estimate = NTB, se = sigmaTrans.NTB, p.value = 2*(1-pnorm(NTB/sigma.NTB)),
    pTrans.value = 2*(1-pnorm(atanh(NTB)/sigmaTrans.NTB)))
```

1st order H-decomposition (rescaled)

R output

```
  estimate           se      p.value pTrans.value
0.14787764   0.09096860   0.09652625   0.10150573
```

Inference based on permutations

Using resampling:

1. Permute the group labels (possibly within strata) in P different ways.
2. For each estimate the Net Treatment Benefit ($\Delta^{\mathcal{P}(1)},\ldots,\Delta^{\mathcal{P}(P)}$)
3. Evaluate the frequency of a more extreme result

$$p^{\mathcal{P}} = \frac{1}{1+P}\left\{1 + \sum_{p=1}^{P} \mathbb{1}_{\left|\Delta^{\mathcal{P}(p)}\right| \geq |\Delta|}\right\}$$

Using an analytic formula (*Anderson & Verbeeck, Mathematics, 2023*)

- matches Wilcoxon's test p-value (no ties)
- assumes normally distributed test statistic
- currently only implemented for Gehan's score

Equivalence GPC and Wilcoxon rank sum test

(without continuity correction)

R code

```
> eBT.perm <- BuyseTest(treatment ~ cont(score), data = data,
                        method.inference = "varexact permutation")
> model.tables(eBT.perm)
```

R output

```
  endpoint total favorable unfavorable neutral uninf  Delta   p.value
1    score   100     53.67       46.33       0     0 0.0734 0.3698664
```

R code

```
> wilcox.test(score ~ treatment, data = data, correct = FALSE)$p.value
```

R output

```
0.3698664
```

Argument method.inference in Buyse Test

Possible values:

- `"none"`
- `"u statistic"` (default)
- `"varexact permutation"`
- `"permutation"`
- `"studentized permutation"`
- `"bootstrap"`
- `"studentized bootstrap"`

Additional arguments

`transformation` (T/F): in follow-up methods (e.g. summary)

`n.resampling`: number of samples

`strata.resampling`: stratified resampling

`cpus`: parallel evaluation

`seed`: for reproducibility

Change default behavior

R code

```
BuyseTest.options(method.inference = "permutation", n.resampling = 1000,
                  statistic = "winRatio")
```

Examples revisited

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

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

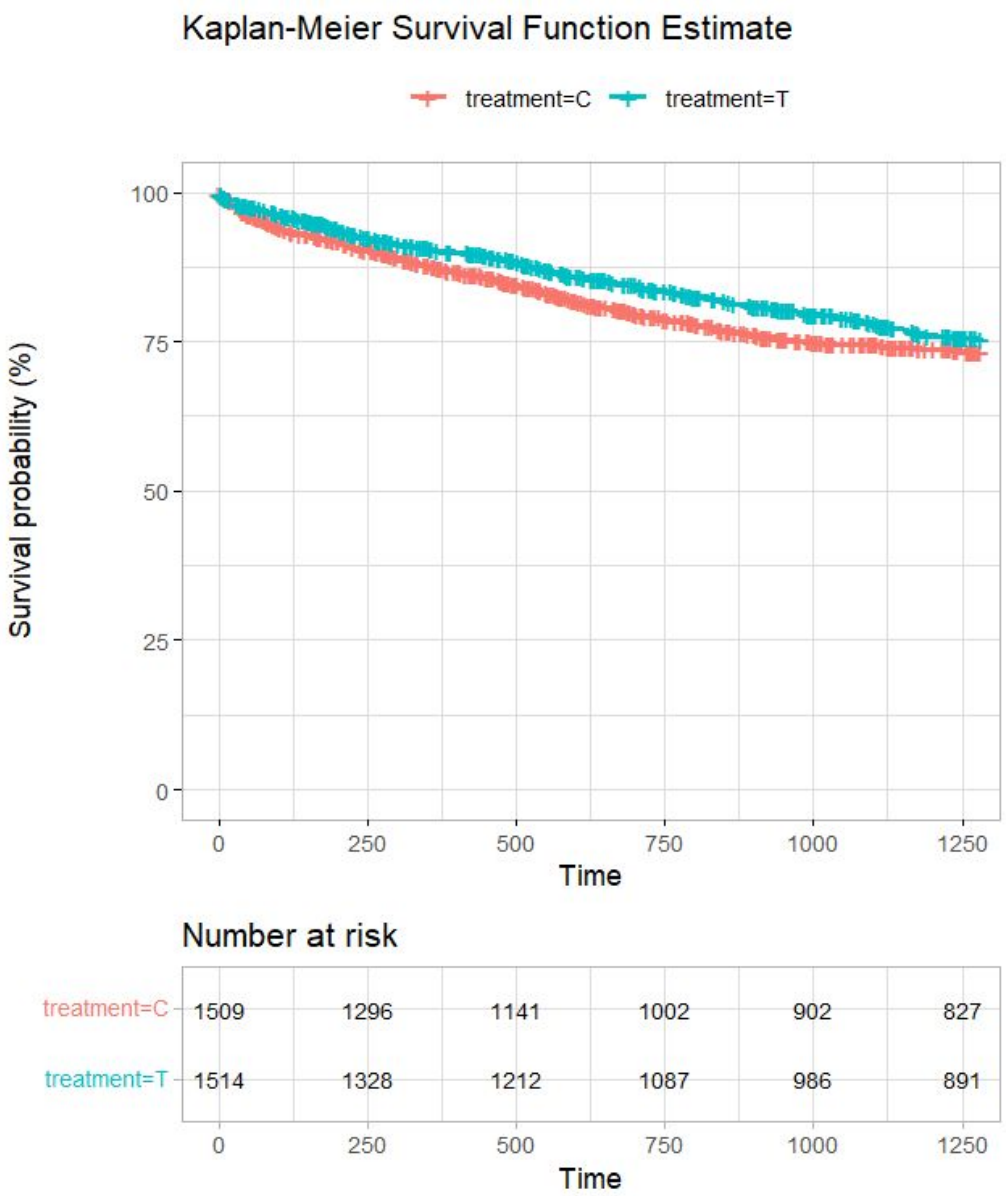
```
> head(charm)
  treatment Mortality statusMortality Hospitalization statusHospitalization Composite statusComposite
1         C  3.891039               1        3.891039                     0  3.891039               1
2         C  3.929701               1        3.929701                     0  3.929701               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

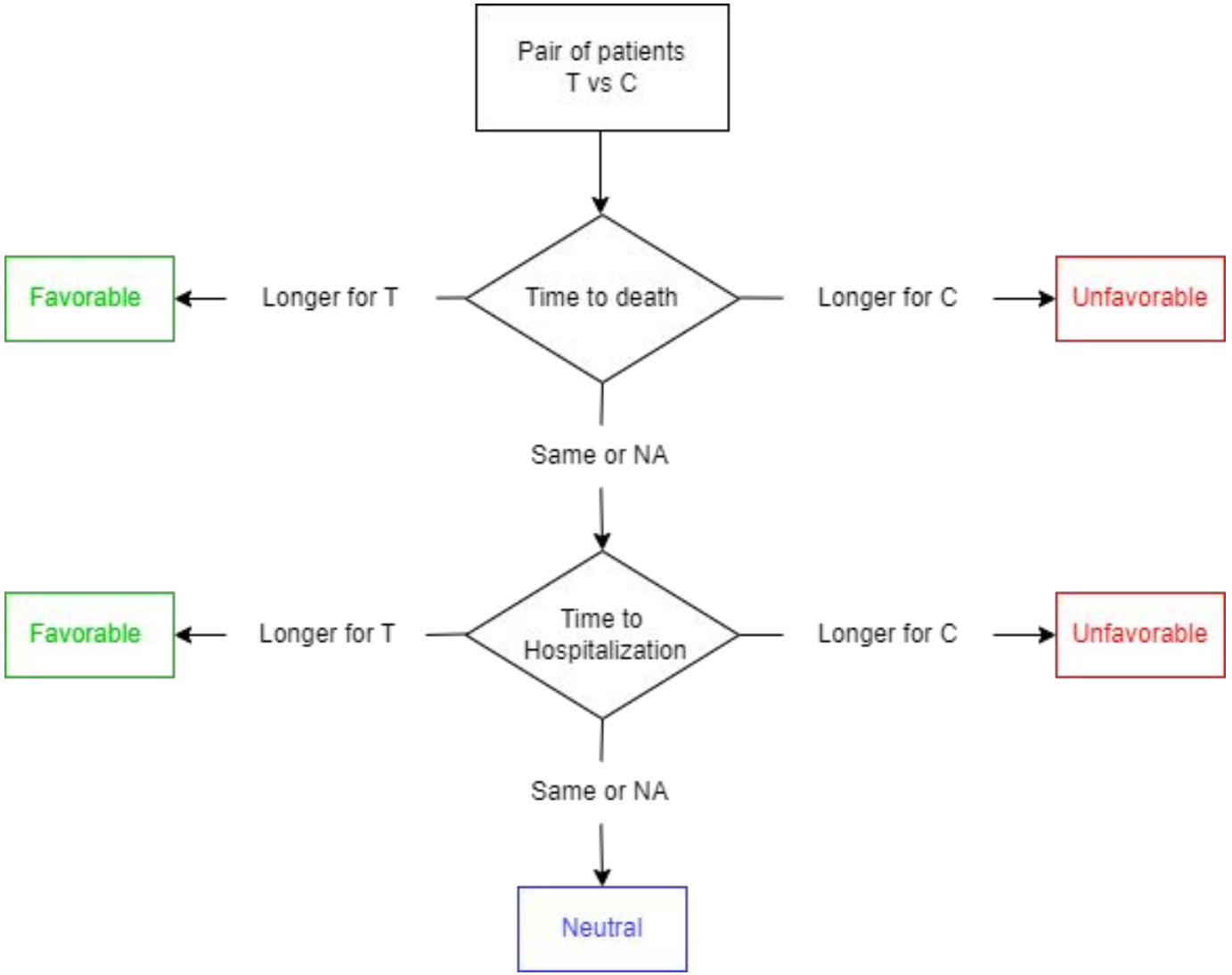
Candesartan	Placebo
92 (54%)	90 (53%)
241 (100%)	276 (100%)

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)

```
> BT14_CHARM <- BuyseTest(treatment~tte(Mortality,statusMortality, threshold=14) +
+                           tte(Hospitalization,statusHospitalization, threshold=14),
+                         data=CHARM, scoring.rule = "Gehan", trace=0)
> summary(BT14_CHARM)
       Generalized pairwise comparisons with 2 prioritized endpoints

 - statistic       : net treatment 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%]  p.value
      Mortality        14   100.00         9.47           8.95       0.03    81.55 0.0052 0.0052   [-0.0147;0.025] 0.609892
 Hospitalization       14    81.58        10.51           7.94       0.04    63.09 0.0257 0.0308   [0.0033;0.0584] 0.028366 *
```

2. Benefit-Risk assessment revisited

- GPC analysis takes multiple prioritized outcomes into account:
  1. Survival gain of at least 6 months
  2. Worse side effect reduction by at least 2 grades
  3. Survival gain of at least 1 month
  4. Any reduction in worse side effect

- Such an analysis is more clinical relevant *and* more powerful
  - Illustration using a simulated dataset resembling the prodige trial

R code

```
> data("prodige", package = "BuyseTest")
> head(prodige)
```

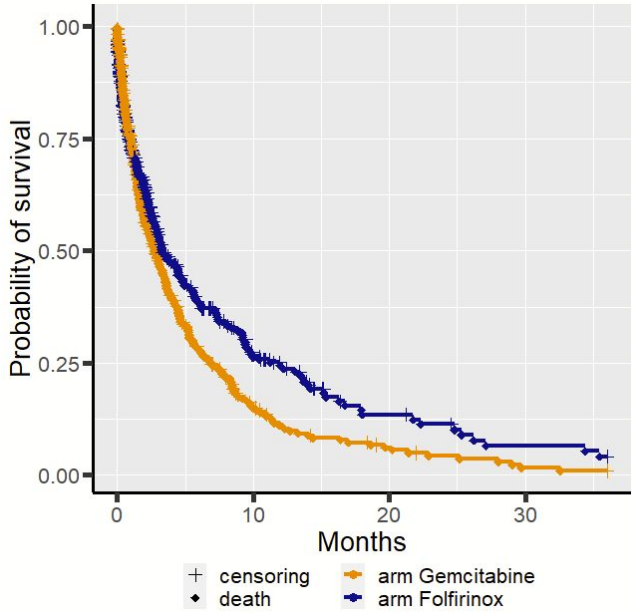
R output

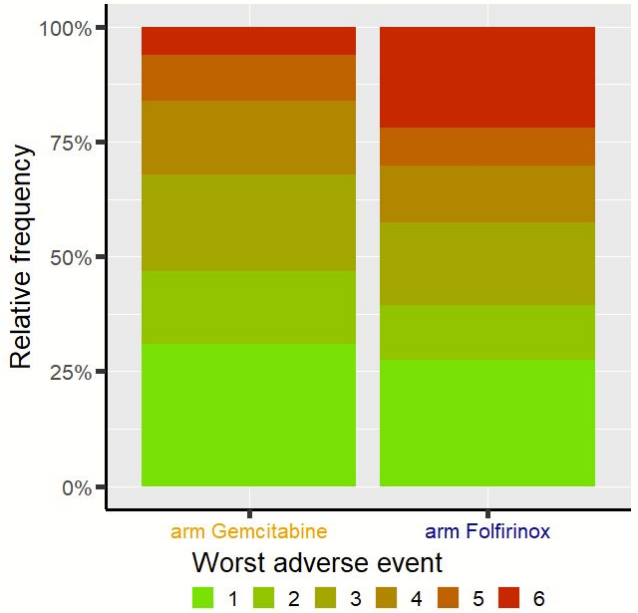
```
    id treatment     OS statusOS    PFS statusPFS toxicity    sex
<num>    <fctr>  <num>    <num>  <num>     <num>    <num> <fctr>
1:     1         C 0.0349        1 0.0349         0        1      F
2:     2         C 2.2790        0 2.2052         1        4      F
3:     3         C 0.2008        1 0.2008         0        1      M
4:     4         C 0.3418        1 0.3418         0        1      F
```

Analyses

- Marginal
  - Hazard ratio: 0.778 [0.658,0.920]
  - Probability of toxicity ≥ 3

    53.3% vs 60.6% (p = 0.033)

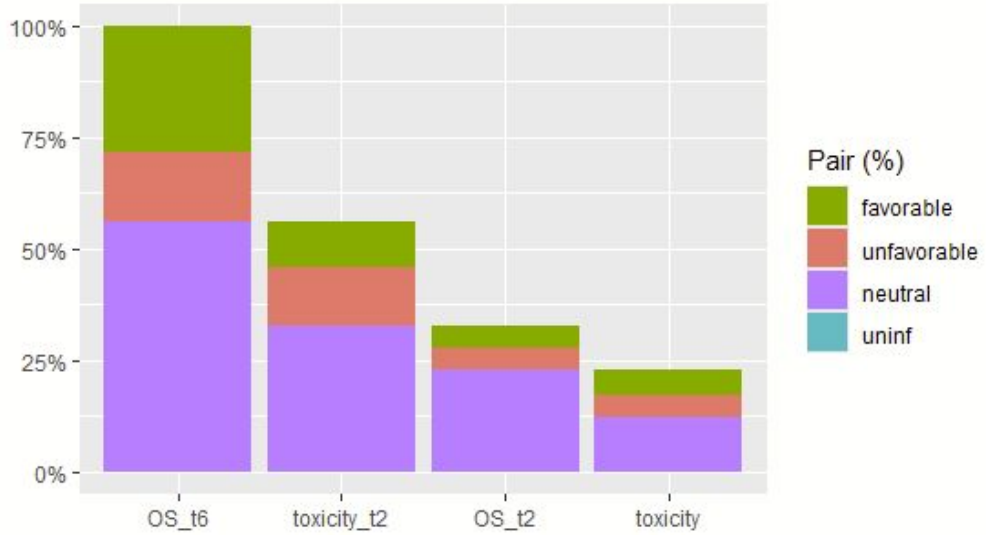




- Joint analysis using GPC

R code

```
> e.BR <- BuyseTest(treatment ~ tte(OS, statusOS, threshold = 6)
                  + cont(toxicity, operator = "<0", threshold = 2)
                  + tte(OS, statusOS, threshold = 1)
                  + cont(toxicity, operator = "<0"),
                      data = prodige)
> plot(e.BR)
```

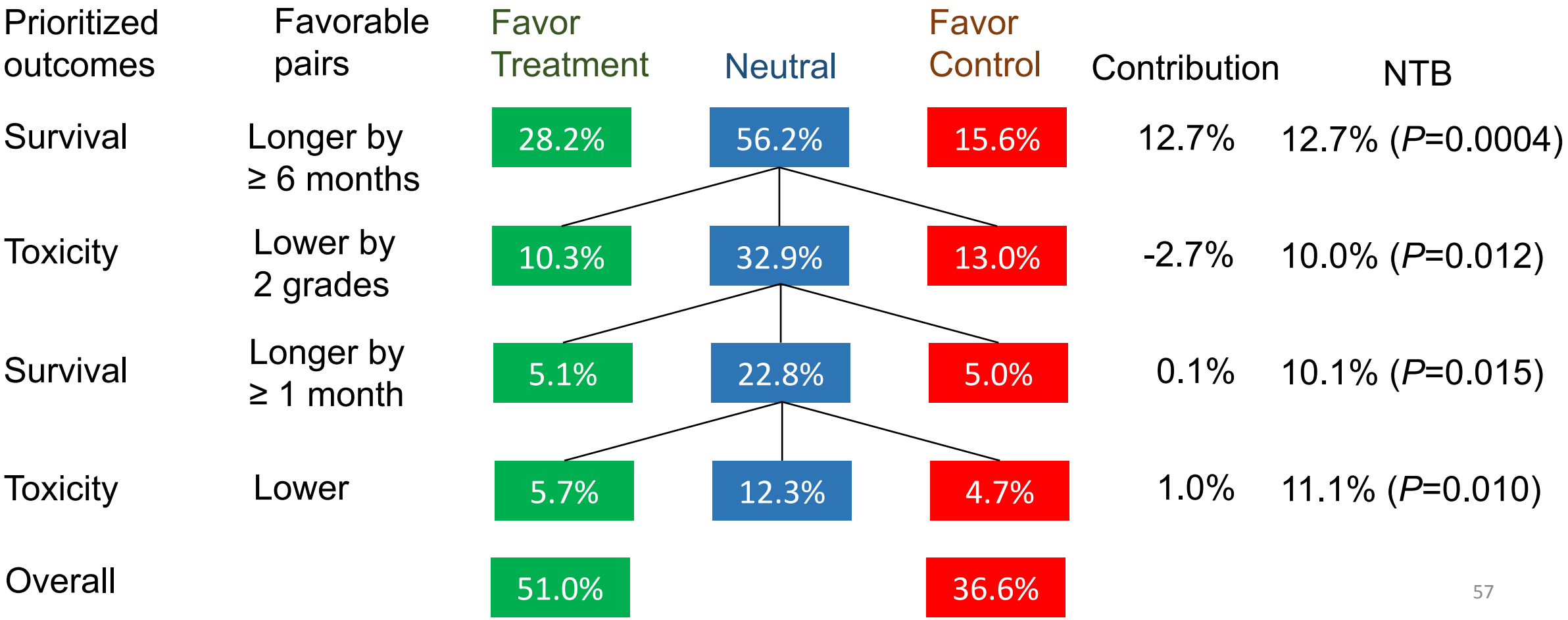


Joint analysis using GPC

R code

```
> summary(e.BR)
```

(simplified output)



GPC analysis of multivariate outcome

- 11.1% = (28.2% - 15.6%) + (10.3% - 13.0%) + (5.1% - 5.0%) + (5.7% - 4.7%)

  $NTB = NTB_1 + NTB_2 + NTB_3 + NTB_4$

  - additive contributions of outcomes: facilitates interpretation

  ⚠ NTB_2 , NTB_3 , NTB_4 do not reflect marginal effects

  they depends on previous outcomes (probability of neutral pair)

- Win Ratio = 49.4% / 38.3% = 1.29
  - no additive or multiplicative decomposition
  - difficult to understand the contribution of each outcome

Sensitivity analysis

Sensitivity of the result to the choice of the threshold(s)

R code

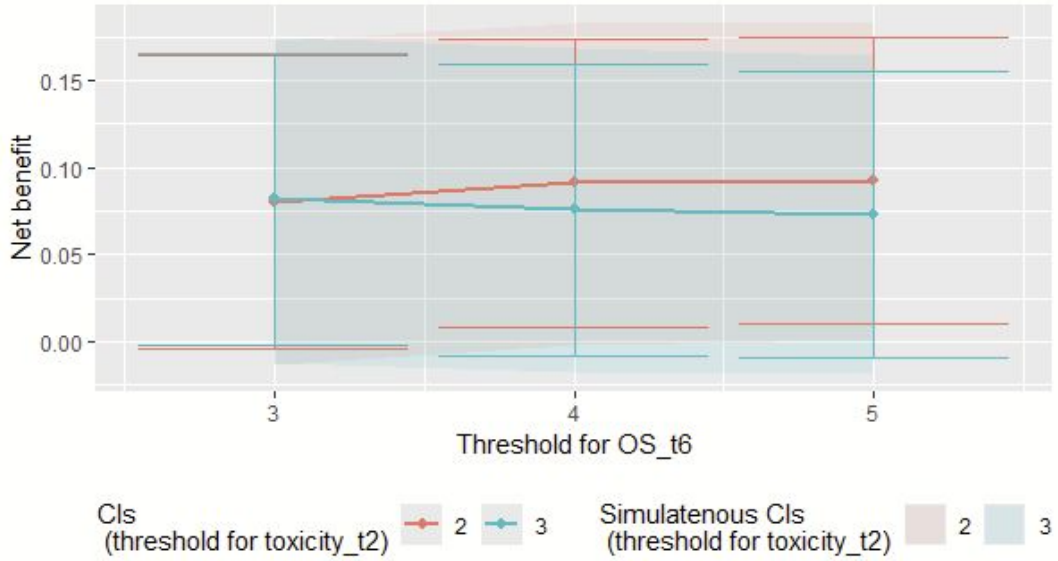
```
> M.threshold <- cbind(OS_t6 = c(3:5,3:5),
                       toxicity_t2 = c(2,2,2,3,3,3),
                       OS_t2 = 1,
                       toxicity = 0)
> M.threshold
```

R output

```
     OS_t6 toxicity_t2 OS_t2 toxicity
[1,]     3           2     1        0
[2,]     4           2     1        0
[3,]     5           2     1        0
[4,]     3           3     1        0
[5,]     4           3     1        0
[6,]     5           3     1        0
```

R code

```
> eBR.Se <- sensitivity(e.BR, band = TRUE,
                        threshold = M.threshold)
> plot(eBR.Se)
```

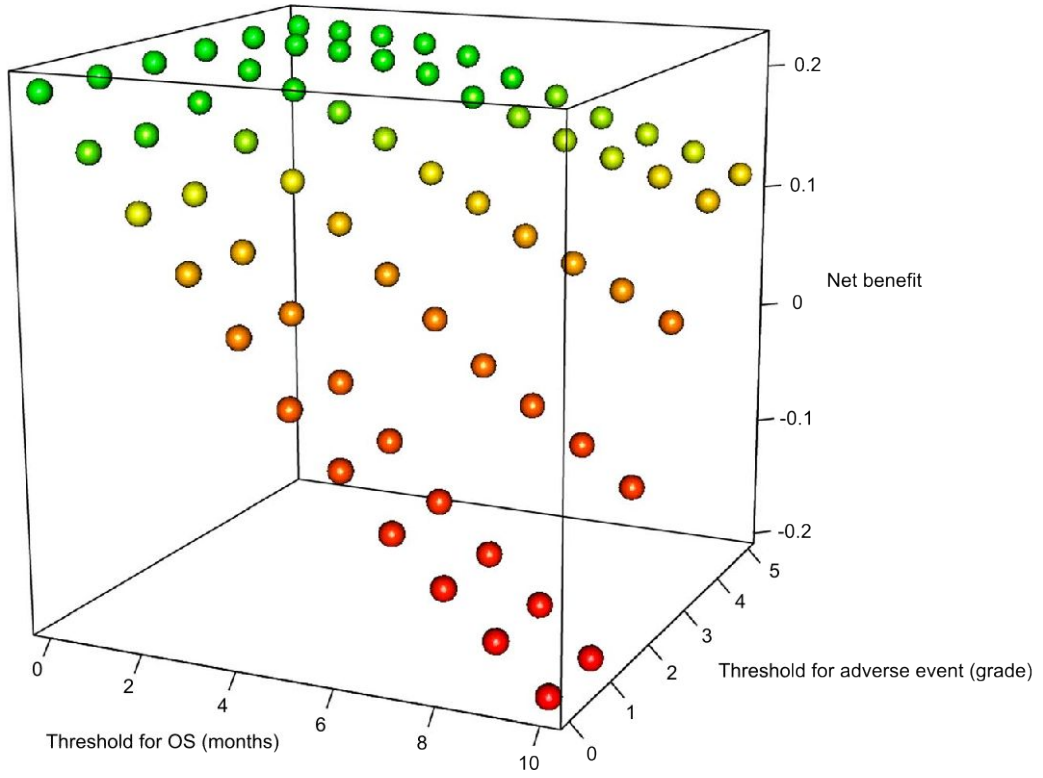


in other applications, the choice of the threshold can have substantial impact on the results

Sensitivity analysis

The Benefit-Risk Balance of Nab-Paclitaxel in Metastatic Pancreatic Adenocarcinoma

Julien Péron, MD, PhD,† Joris Giai, MD,* Delphine Maucort-Boulch, MD,* and Marc Buyse, ScD‡*



3. EB rare disease trial revisited

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

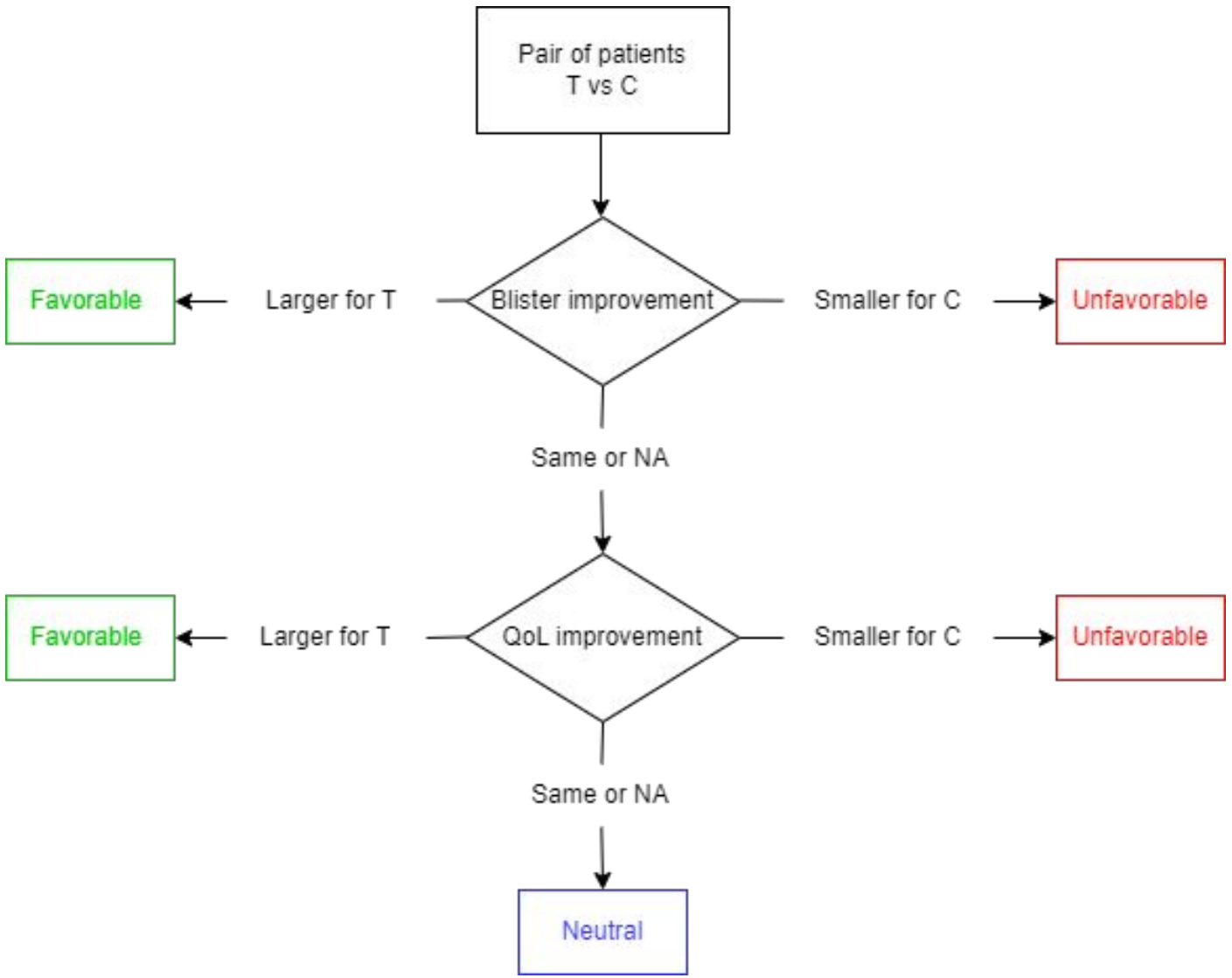
- Patient-centric analysis: blister count and change in QoL at week 4

- Uncertainty in blister counts

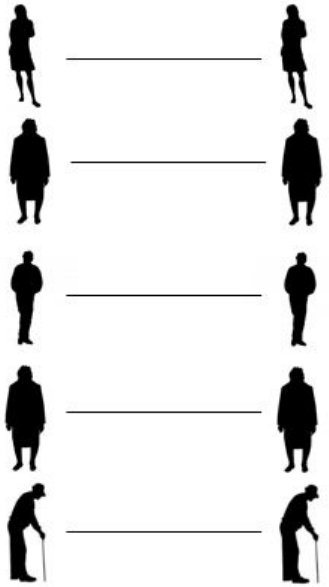
EB revisited:

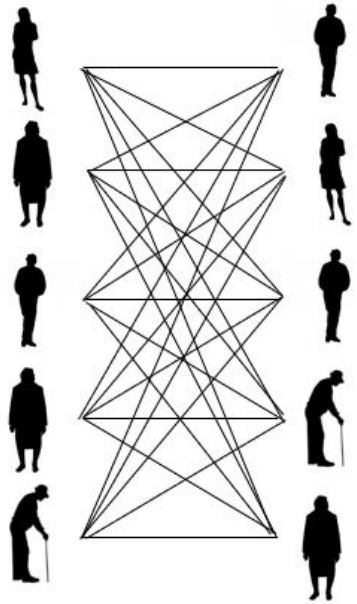
```
> head(EB)
    Id Time Group StdDiffCount Bin DiffQoL period
1 1001   t4     V    0.6666667   1       2      1
2 1001  t12     P    0.0000000   0       0      2
3 1002   t4     P   -0.2500000   0      -1      1
4 1002  t12     V   -4.0000000   0       0      2
5 1004   t4     V    0.5454545   1 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

EB revisited: Patient-centric outcome



Matched versus unmatched GPC





$$\Delta_m = P\left(Y_i^E > Y_i^C\right) - P\left(Y_i^E < Y_i^C\right)$$

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

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

requires at least 15-20 (paired) subjects

Konietschke et al. *Electron J Stat* (2012)
Fay et al. *Stat Med.* (2018)
Matsouaka SMMR (2022)
Verbeeck et al. OJRD. (2023)

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

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

         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 all possible permutations
   - cpus  : 1

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 0.0701057
  DiffQoL    45.33           32           6.22       5.33     1.78 0.2578 0.5911 0.0051302
```

EB revisited: less evidence for count outcome

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

         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 all possible permutations
   - cpus  : 1

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 0.069625
      DiffQoL     8.89         6.22           0.44       1.78     0.44 0.0578 0.4267 0.040017
```

EB revisited: accounting for blister uncertainty

```
> print(BuyseTest(Group~c(StdDiffCount, threshold=0.2)+c(DiffQoL), data=EB,method.inference="varexact-permutation"))

         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 all possible permutations
   - cpus  : 1

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 0.059927
      DiffQoL                  24        15.11           2.67       3.56     2.67 0.1244 0.4933 0.016440
```

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

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

```
> 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

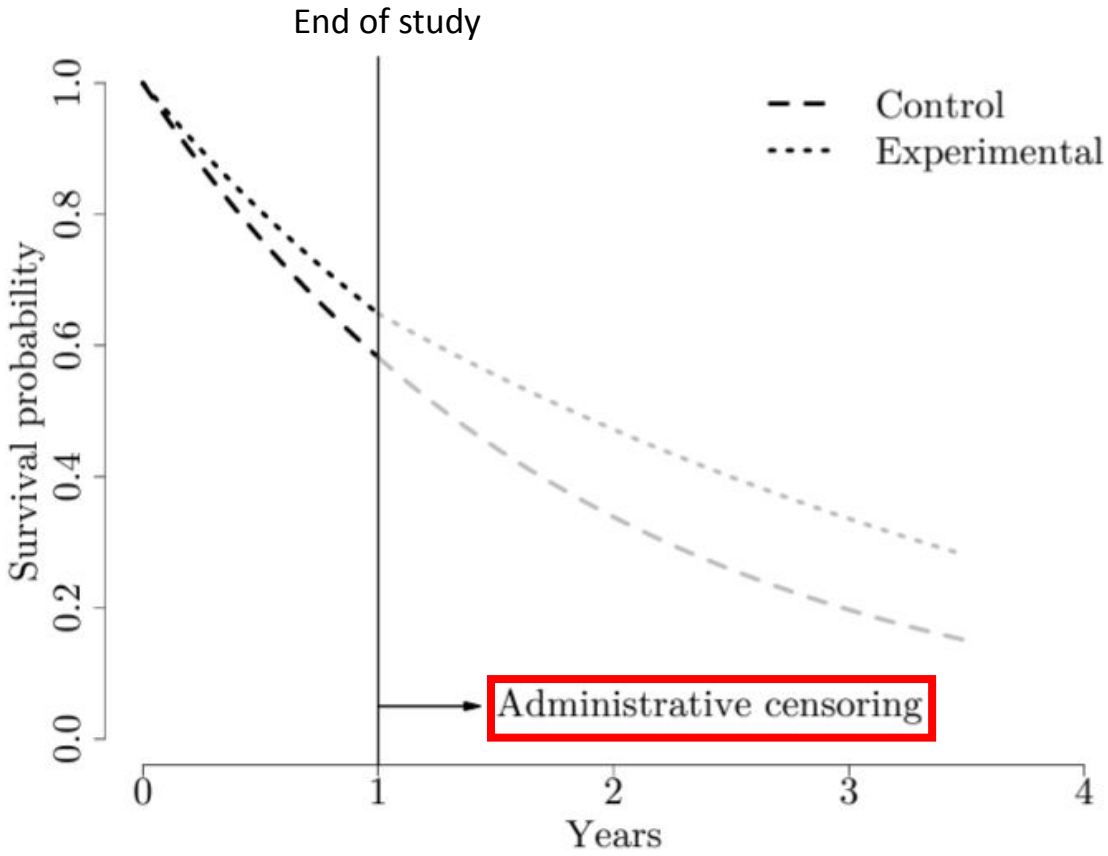
```
  p.value
0.0701057
0.0051302
```

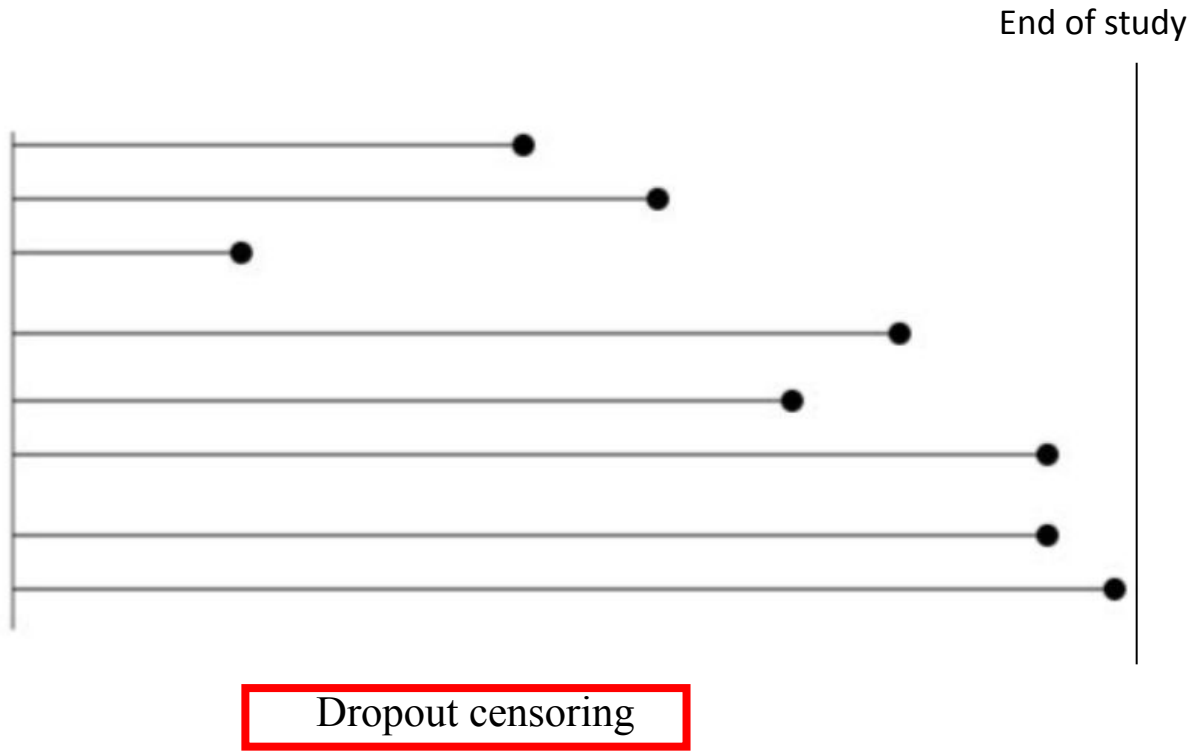
Questions?

Break

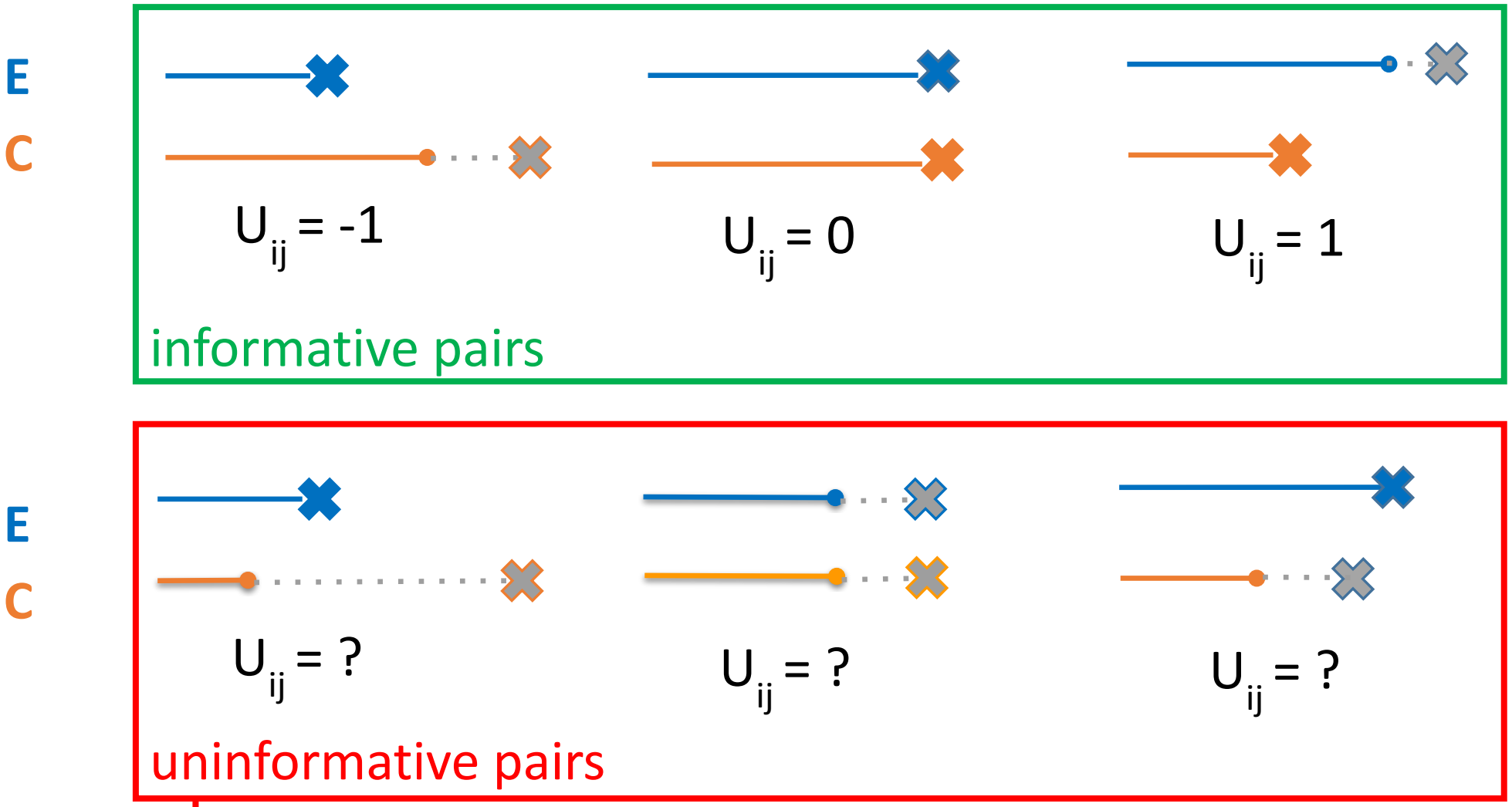
Advanced Topics

1. Censoring

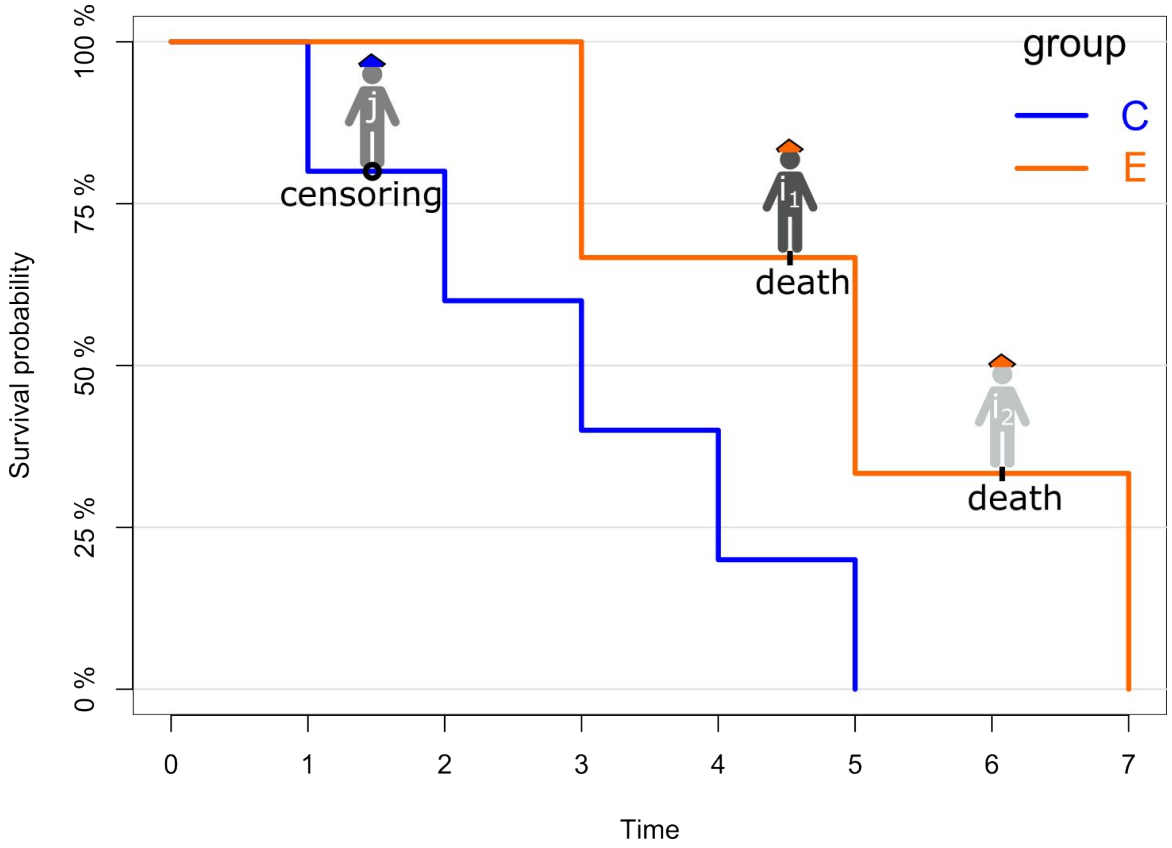

End of study
Control
Experimental
Survival probability
0.0
0.2
0.4
0.6
0.8
1.0
Administrative censoring
0
1
2
3
4
Years


End of study
Dropout censoring

Gehan scoring rule (recap')



Peron scoring rule (example)



$$\left(\widetilde{Y}^E_{i_1}, \widetilde{Y}^E_{i_2}, \widetilde{Y}^C_j, \Omega^E_{i_1}, \Omega^E_{i_2}, \Omega^C_j\right) = (4.7, 6.1, 1.5, 1, 1, 0)$$

time to event (right-censored)

event type indicators

$$\mathbb{P}\left[Y^E_i > Y^C_j | \widetilde{Y}^E_i, \Omega^E_i, \widetilde{Y}^C_j, \Omega^C_j\right] = \begin{cases} 0.75 \text{ for } i = i_1 \\ 1 \text{ for } i = i_2 \end{cases}$$

Only 1 out 4 individuals in the other group will survive up to the time of the observed event.

No survivor in the other group at the time of the observed event.

Peron scoring rule (formula)

$$U_{ij} = \mathbb{P}\left[Y_i > Y_j + \tau | \widetilde{y}_i, \omega_i, \widetilde{y}_j, \omega_j\right] - \mathbb{P}\left[Y_j > Y_i + \tau | \widetilde{y}_i, \omega_i, \widetilde{y}_j, \omega_j\right]$$

pairwise score

(ω_i, ω_j)	$\tilde{y}_i - \tilde{y}_j \leq -\tau$	$\lvert\tilde{y}_i - \tilde{y}_j\rvert < \tau$	$\tilde{y}_i - \tilde{y}_j \geq \tau$
$(1,\ 1)$	-1	0	1
$(0,\ 1)$	$\frac{S(\tilde{y}_j+\tau)+S(\tilde{y}_j-\tau)}{S(\tilde{y}_i)} - 1$	$\frac{S(\tilde{y}_j+\tau)}{S(\tilde{y}_i)}$	1
$(1,\ 0)$	-1	$-\frac{S(\tilde{y}_i+\tau)}{S(\tilde{y}_j)}$	$1 - \frac{S(\tilde{y}_i+\tau)+S(\tilde{y}_i-\tau)}{S(\tilde{y}_j)}$
$(0,\ 0)$	$\ldots$	$\frac{\int_{\tilde{y}_i}^{\infty} S(t+\tau)dS(t) - \int_{\tilde{y}_j}^{\infty} S(t+\tau)dS(t)}{S(\tilde{y}_i)S(\tilde{y}_j)}$	$\ldots$

event type indicators
- 0 censored
- 1 event

Large right-censored event time for the control arm

'Similar' right-censored event times

Large right-censored event time for the experimental arm

Péron et al. SMMR (2016)

Handling right-censoring with BuyseTest

- Naïve approaches: (BuyseTest: scoring.rule = "Gehan“)
  - uninformative pairs = 0
  - biased but easy to carry out and preserves type 1 error control

Gehan Biometrika (1965)

- Imputation approaches: (BuyseTest: scoring.rule = “Peron”)
  - estimate a probability per pair based on a survival model
  - default model: Kaplan-Meier stratified on treatment arm
  - estimation and uncertainty quantification is more complex

Péron et al. SMMR (2016)

Ozenne et al. SMMR (2021)

- Add-hoc/other approaches: (BuyseTest: correction.uninf=1 or 2)
  - superior alternatives

Péron et al. Biom J (2021)

Using build-in imputation approach

R code

```
e.NTB_Gehan <- BuyseTest(treatment ~ tte(OS, statusOS), scoring.rule = "Gehan",
                         data = prodige, keep.pairScore = TRUE, trace = FALSE)
getPairScore(e.NTB_Gehan)[1:2,]
```

store pairwise score (U_{ij})

memory intensive

R output

```
   index.C index.T favorable unfavorable neutral uninf weight
1:       1     403         1           0       0     0      1
2:       2     403         0           0       0     1      1
```

R code

```
e.NTB_Peron <- BuyseTest(treatment ~ tte(OS, statusOS), scoring.rule = "Peron",
                         data = prodige, keep.pairScore = TRUE, trace = FALSE)
getPairScore(e.NTB_Peron)[1:2,]
```

R output

```
   index.C index.T favorable unfavorable neutral       uninf weight
1:       1     403 1.0000000     0.00000       0 0.0000000000      1
2:       2     403 0.5286551     0.47068       0 0.0006648516      1
```

Several other corrections

See *Deltuvaite-Thomas et al. Biometrical journal (2022)* for an overview

- <u>Naïve approaches</u>:
  - Gehan: uninformative pairs = 0 — *Gehan Biometrika (1965)*
  - Harrell: ignore uninformative pairs — *Harrell et al. JAMA (1982)*

- <u>Imputation approaches</u>: survival model
  - Latta: Kaplan-Meier (common to both arms) — *Latta Biometrika (1977)*
  - Peron: Kaplan-Meier stratified on treatment arm — *Péron et al. SMMR (2016)*
  - Efron: same but constrained to 0 at end of follow-up — *Efron Proc 5th Berkeley Symposium on Math. Stat. and Proba. (1967)*
  - De Backer: use extreme value tail model — *De Backer Pharm Stat. (2023)*

- <u>Weighting approaches</u>: inverse probability of censoring
  - Datta: pairs with censored event weight 0 — *Datta et al. Scandinavian Journal of Statistics (2010)*
  - Dong: 'Gehan'-like alternative — *Dong et al. Journal of Biopharmaceutical Statistics (2020)*

The Kaplan Meier estimators ($S_{KM}(t)$ event of interest, $C_{KM}(t)$ censoring) satisfy:
$S_{KM}(t) \cdot C_{KM}(t)$ = number still at risk at time t+ / sample size
-> similarities between weighting and imputation: Efron/Datta Peron/Dong

(Personal) critic of the other corrections

implemented in BuyseTest

not implemented in BuyseTest

- Naïve approaches:
  - Gehan: uninformative pairs = 0
  - Harrell: ignore uninformative pairs → biased (except PH case)

- Imputation approaches: survival model → argument `model.tte` to provide an adequate survival model
  - Latta: Kaplan-Meier
  - Peron: Kaplan-Meier stratified on treatment arm
  - Efron: same but constrained to 0 at end of follow-up
  - De Backer: use extreme value tail model

- Weighting approaches: inverse probability of censoring
  - Datta: pairs with censored event weight 0 → 'discard' information
  - Dong: 'Gehan'-like alternative → equivalent to imputation approach with un-natural survival models

→ Useful to handle informative censoring if covariates available

→ Unclear validity of the extension to multiple, correlated, endpoints

Using your own imputation approach

R code

```
> e.NTB_Latta <- BuyseTest(treatment ~ tte(OS, statusOS), scoring.rule = "Peron",
                           data = prodige, trace = FALSE,
                           model.tte = prodlim(Hist(OS, statusOS) ~ 1, data = prodige))
```

Latta

`prodlim`: Kaplan Meier estimator (possibly stratified)

`survreg`: parametric survival model (possibly with covariates)

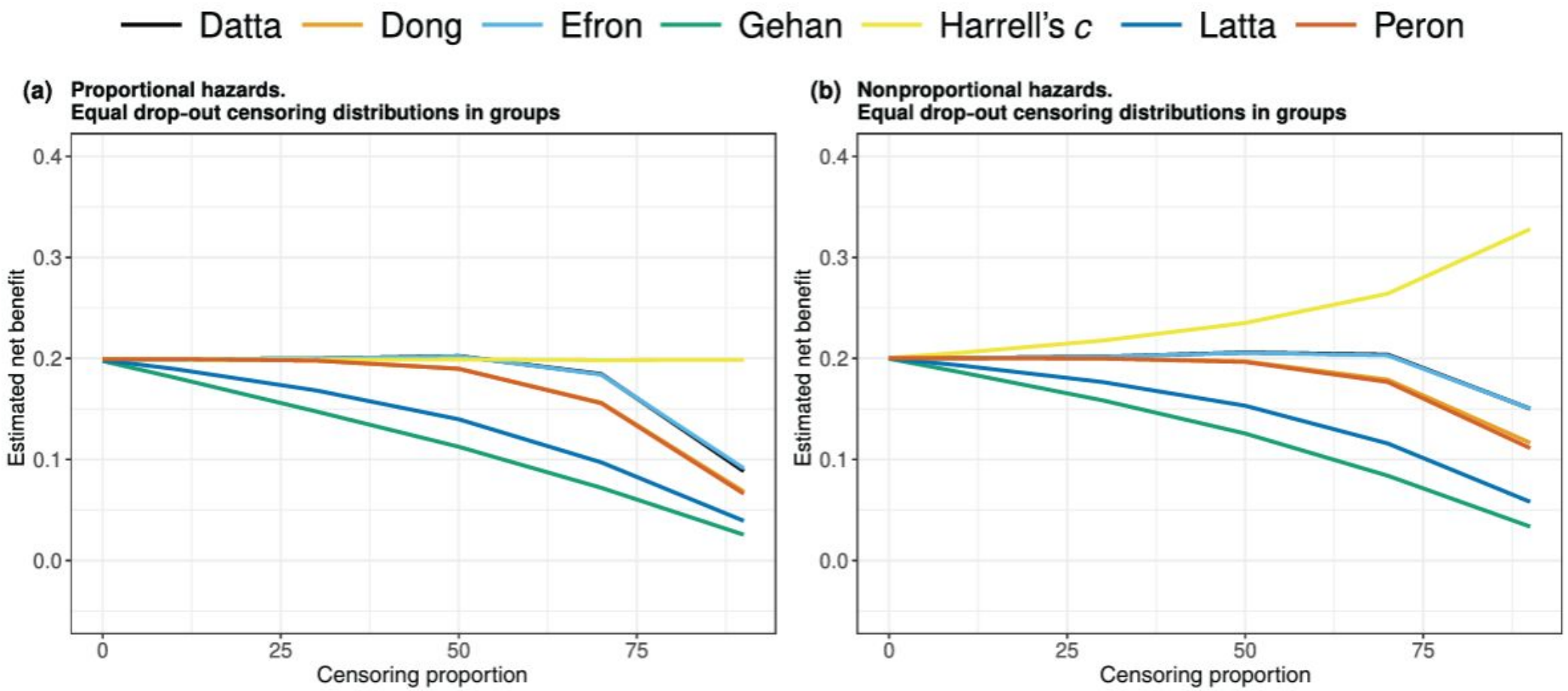
- rely on numerical integration

⚠ Software limitation:

- only implemented for survival models with covariates that are constant within the sub-groups formed by the treatment and strata variables.

Performance in presence of drop-out

- Reasonable for censoring below 50%



Deltuvaite-Thomas et al. Biometrical journal (2022)

What about administrative censoring?

- Real life studies have finite follow-up time
  - The tail of the survival function
    is not (non-parametrically) identifiable
  - May be needed to score some of the pairs!

(ω_i, ω_j)	$\tilde{y}_i - \tilde{y}_j \leq -\tau$	$\lvert\tilde{y}_i - \tilde{y}_j\rvert < \tau$	$\tilde{y}_i - \tilde{y}_j \geq \tau$
$(1,\ 1)$	-1	0	1
$(0,\ 1)$	$\frac{S(\tilde{y}_j+\tau)+S(\tilde{y}_j-\tau)}{S(\tilde{y}_i)} - 1$	$\frac{S(\tilde{y}_j+\tau)}{S(\tilde{y}_i)}$	1
$(1,\ 0)$	-1	$-\frac{S(\tilde{y}_i+\tau)}{S(\tilde{y}_j)}$	$1 - \frac{S(\tilde{y}_i+\tau)+S(\tilde{y}_i-\tau)}{S(\tilde{y}_j)}$
$(0,\ 0)$	$\dots$	$\frac{\int_{\tilde{y}_i}^{\infty} S(t+\tau)dS(t) - \int_{\tilde{y}_j}^{\infty} S(t+\tau)dS(t)}{S(\tilde{y}_i)S(\tilde{y}_j)}$	$\dots$

- Possible remedies:
  - make parametric assumptions (Efron, De Backer, Harrel, …)
  - switch to an easier estimand: restricted NTB to time Γ (here Γ =24 months)

    rNTB = $\mathbb{P}[\min(X,\Gamma)>Y] - \mathbb{P}[\min(Y,\Gamma)>X]$

Piffoux et al. J Clin Epidemiol (2024)

R code

```
> e.NTB_restricted <- BuyseTest(treatment ~ tte(OS, statusOS, restriction = 24),
                                scoring.rule = "Peron", data = prodige)
```

2. Covariate adjustment

1. Marginal treatment effect

- Expected improvement in population
- Simplicity of communication; reduced complexity for combining results; decision making on population level

$$\Delta_\tau := P\left(Y_i^E \succ_\tau Y_j^C | A^E = 1, A^C = 0\right) - P\left(Y_i^E \prec_\tau Y_j^C | A^E = 1, A^C = 0\right)$$

2. Conditional treatment effect

- Expected improvement conditionally on characteristics shared by individuals
- Closer to individual-level effect

$$\Delta_\tau(X^E, X^C) := P\left(Y_i^E \succ_\tau Y_j^C | A^E = 1, A^C = 0, X^E, X^C\right) - P\left(Y_i^E \prec_\tau Y_j^C | A^E = 1, A^C = 0, X^E, X^C\right)$$

Collapsibility

- Marginal effect = (weighted) sum of conditional effects in subgroups or individuals

- Many treatment effect measures are non-collapsible (OR, HR,..)

- GPC treatment effect measures (PI, NTB, WR, SO) are non-collapsible

  -> Conditional effects may differ from marginal effects

  -> Important to ascertain interest in marginal or conditional effect a priori

Strategies for covariate adjustment in GPC

1. **Stratification**

   Conditional treatment effects

2. **Non-parametric adjustment**

   Marginal treatment effect

3. **Semi-parametric models: Generalized PI models (GPIM)**

   Marginal and conditional effects

Stratification in GPC

Recall the U-statistic

$$\widehat{\Delta} = \frac{1}{n^E n^C} \sum_{i=1}^{n^E} \sum_{j=1}^{n^C} U_{ij}$$

With *K* strata, this becomes

$$\widehat{\Delta}^s = \sum_{k=1}^{K} w_k \widehat{\Delta}^k$$

$$\widehat{\Delta}^k = \frac{1}{n_k^E n_k^C} \sum_{i=1}^{n_k^E} \sum_{j=1}^{n_k^C} U_{ij}$$

This U-statistic is a conditional estimate of the NTB, given the strata

Buyse. Stat Med (2010)
*Dong et al. Pharmaceutical Statistics (*2023)

Choice of weights

- **Cochran-Mantel-Haenszel (CMH) weights**: weights proportional to the ratio of the strata pairs and the observations per strata = best (pool.strata=”CMH”)

$$w_k \propto \frac{n_k^E n_k^C}{n_k^E + n_k^C}$$

- Weights proportional to the strata pairs (pool.strata=”buyse”)

$$w_k \propto n_k^E n_k^C$$

- Equal weights (pool.strata=”equal”)

$$w_k = {}^1/_K$$

- Inverse variance weights (pool.strata=”var-netBenefit”)

$$w_k \propto {}^1/_{\hat{\sigma}_k^2}$$

Buyse. Stat Med (2010)
*Dong et al. Pharmaceutical Statistics (*2023)

Revisit Prodige trial with sex stratification

```
> s.BR <- BuyseTest(treatment ~ tte(OS, statusOS, threshold = 6) + cont(toxicity, operator = "<0", threshold = 2) +
+                   tte(OS, statusOS, threshold = 2) + cont(toxicity, operator = "<0") + strata(sex),
+                   , pool.strata = "CMH", data = prodige)
```

endpoint	threshold	strata	total(%)	favorable(%)	unfavorable(%)	neutral(%)	uninf(%)	delta	Delta CI [2.5% ; 97.5%]	p.value	
OS	6	global	100.00	28.22	15.64	56.04	0.10	0.126	0.126 [0.056;0.195]	0.00048	***
		M	49.04	13.79	6.84	28.33	0.08	0.142			
		F	50.96	14.43	8.80	27.71	0.02	0.111			
toxicity	2	global	56.14	10.30	13.07	32.77	0.00	-0.028	0.098 [0.02;0.176]	0.01410	*
		M	28.41	5.52	6.28	16.60	0.00	-0.016			
		F	27.73	4.77	6.78	16.17	0.00	-0.039			
OS	2	global	32.77	4.99	5.02	22.72	0.05	0.000	0.098 [0.016;0.178]	0.01856	*
		M	16.60	2.22	2.46	11.87	0.05	-0.005			
		F	16.17	2.77	2.56	10.85	0.00	0.004			
toxicity		global	22.76	5.71	4.71	12.35	0.00	0.010	0.108 [0.023;0.192]	0.01312	*
		M	11.92	2.83	2.39	6.70	0.00	0.009			
		F	10.85	2.88	2.31	5.65	0.00	0.011			

Revisit Prodige trial with sex stratification

Explore size of strata and number of pairs

```
> nobs(s.BR, strata = TRUE)
    C   T pairs
M 190 218 41420
F 212 203 43036
```

Explore NTB per strata and per endpoint

```
> confint(s.BR, strata = TRUE)
                estimate         se     lower.ci  upper.ci null     p.value
OS_t6.M       0.14163397 0.04738957  0.047772987 0.2330156    0 0.003192326
OS_t6.F       0.11055413 0.05319843  0.005450657 0.2132417    0 0.039286663
toxicity_t2.M 0.12608947 0.05325181  0.020703383 0.2287044    0 0.019148860
toxicity_t2.F 0.07116027 0.05904555 -0.045004977 0.1854268    0 0.229711152
OS_t2.M       0.12124870 0.05565770  0.011132977 0.2284589    0 0.031001199
OS_t2.F       0.07528238 0.06110224 -0.044985322 0.1933990    0 0.219667852
toxicity.M    0.13014476 0.05847593  0.014300729 0.2425413    0 0.027778835
toxicity.F    0.08636752 0.06352251 -0.038834727 0.2088999    0 0.176098720
```

Non-parametric adjustment

~Augmented estimator, with a term that depends on the covariates and their link with the endpoints

$$\widehat{\Delta}^{adj} = \widehat{\Delta} - \widehat{V}'_{\boldsymbol{YX}}\widehat{V}^{-1}_{\boldsymbol{XX}}d_{\boldsymbol{X}}$$

$$\widehat{V}_{\boldsymbol{Y}X_h} = \frac{1}{n^E}\widehat{\mathbb{C}}ov(\widehat{\Delta}^E_{\boldsymbol{\tau},i}, \widehat{\Delta}^E_{i,X_h}) + \frac{1}{n^C}\widehat{\mathbb{C}}ov(\widehat{\Delta}^C_{\boldsymbol{\tau},j}, \widehat{\Delta}^C_{j,X_h})$$

$$\widehat{V}_{X_hX_{h\star}} = \frac{1}{n^E}\widehat{\mathbb{C}}ov(\widehat{\Delta}^E_{i,X_h}, \widehat{\Delta}^E_{i,X_{h\star}}) + \frac{1}{n^C}\widehat{\mathbb{C}}ov(\widehat{\Delta}^C_{j,X_h}, \widehat{\Delta}^C_{j,X_{h\star}})$$

$$d_{\boldsymbol{X}} = (\overline{X}^E_1 - \overline{X}^C_1, \ldots, \overline{X}^E_p - \overline{X}^C_p)'$$

This U-statistic is a marginal estimate of the NTB

Koch et al. Stat Med (1998)

Alternative: standardization

For non-continuous baseline covariates

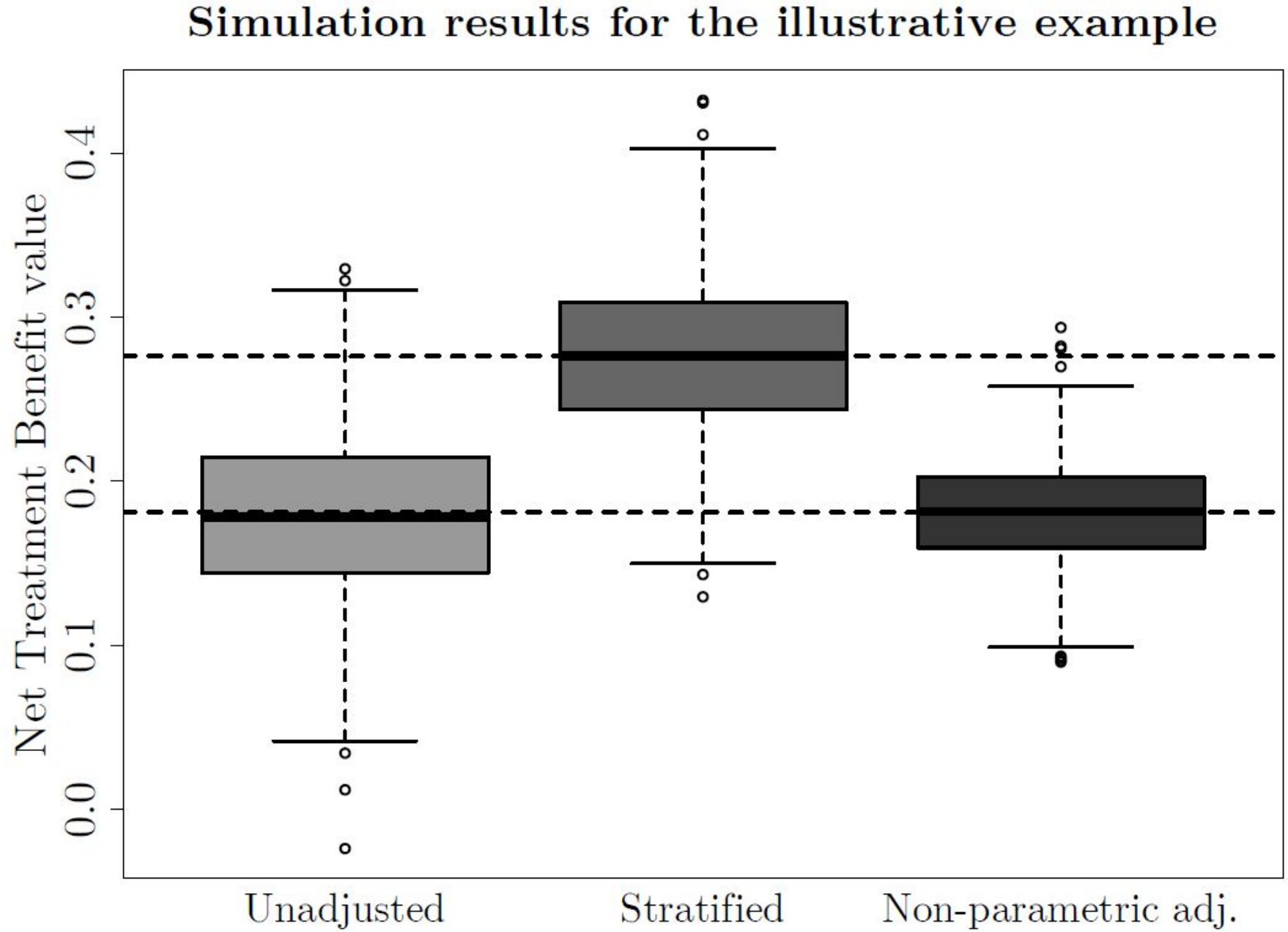
$$\widehat{\Delta}_{\boldsymbol{\tau}}^{std.} = \frac{1}{n^E n^C} \sum_{i=1}^{n^E} \sum_{j=1}^{n^C} w_{\boldsymbol{x}_i, \boldsymbol{x}_j} \{ \mathbb{1}(\boldsymbol{Y}_i^E \succ_{\boldsymbol{\tau}} \boldsymbol{Y}_j^C) - \mathbb{1}(\boldsymbol{Y}_j^C \succ_{\boldsymbol{\tau}} \boldsymbol{Y}_i^E) \},$$

where $w_{\boldsymbol{x}_i, \boldsymbol{x}_j}$ is a weight given to the comparison of individuals i and j, taking value:

$$w_{\boldsymbol{x}_i, \boldsymbol{x}_j} = \frac{n^E n^C}{n^2} \frac{n_{\boldsymbol{x}_i} n_{\boldsymbol{x}_j}}{n_{\boldsymbol{x}_i}^E n_{\boldsymbol{x}_j}^C},$$

for $n_{\boldsymbol{x}_i}$ the total number of individuals sharing the same covariate values as individual i from the experimental arm (and similarly for $n_{\boldsymbol{x}_j}$), and $n_{\boldsymbol{x}_i}^E$ the number of individuals in the experimental arm sharing the same covariate value as individual i (and similarly for $n_{\boldsymbol{x}_j}^C$).

The average absolute difference between the non-parametric adjustment and standardization is of the order 10^{-4} .


Simulation results for the illustrative example
Net Treatment Benefit value
0.0
0.1
0.2
0.3
0.4
Unadjusted
Stratified
Non-parametric adj.

GPIM : Conditional Models

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

$$logit\left(P\left(Y_i^E \succ_\tau Y_i^C | X^E = X^C\right)\right) = \beta_0 + \beta_A(A^E - A^C) + \beta_X'(X^E - X^C)$$

$$\text{expit}(\beta_A) = conditional\ PI$$

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 '*' 0.0
```

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:

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

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

Leveraging small sample GEE -corrections

$$g\left(P\left(Y_i^E \succ_\tau Y_j^C | A^E, A^C, X^E, X^C\right)\right) = \beta_0 + \beta_A(A^E - A^C) + \beta_X'(X^E - X^C)$$

$$V_{\mathrm{LZ}}^{\mathrm{GEE}} = \left(\sum_{k=1}^{K} D_k' V_k^{-1} D\right)^{-1} M_{\mathrm{LZ}} \left(\sum_{k=1}^{K} D_k' V_k^{-1} D\right)^{-1}$$

p	α_1	N	PIM	BR-AJEL	AJEL	MBN	Pan	WL	GST	KC	MD	MK	FG
4	0	14	21.10	18.40	9.88	5.41	21.41	2.08	2.18	4.89	4.16	3.85	11.85
		16	18.01	12.00	9.16	4.98	17.90	2.95	2.75	4.88	3.36	3.76	11.70
		20	11.68	6.29	6.40	3.76	13.40	3.35	4.57	4.47	3.76	3.96	9.24
		24	11.67	5.94	7.24	4.43	11.57	4.53	4.73	5.03	4.63	5.13	9.86
		30	9.53	5.62	6.52	5.12	10.03	5.42	5.82	4.41	4.61	5.02	8.12
	0.5	14	29.92	27.59	17.65	9.62	29.81	4.97	4.12	9.94	7.61	6.77	16.91
		16	28.41	21.23	18.77	11.69	29.23	6.46	7.59	11.08	8.82	8.92	20.10
		20	27.76	17.73	19.35	12.36	28.88	10.03	12.26	12.77	10.84	10.94	22.49
		24	28.51	17.47	20.78	17.17	30.72	15.26	17.87	16.87	16.67	18.07	26.61
		30	34.71	22.74	25.86	21.43	35.51	21.03	22.23	21.03	19.92	21.63	30.68

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

Example of a non-inferiority trial

De Backer, Clin Trials (2024)

Aim:

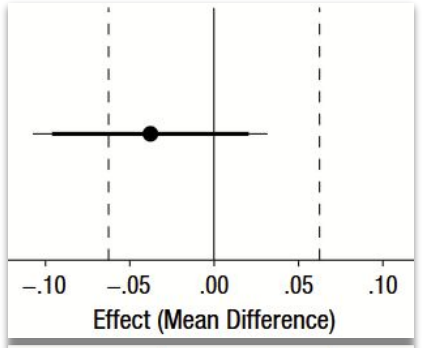
- Obtain approval for a as effective but safer drug
  e.g. Acute Promyelocytic Leukemia: reduced dose treatment vs. full dose

Traditionally

- Primary endpoint: Event-Free Survival (EFS) at 2 years
  -> Expensive: large sample size required to meet a 'narrow' non-inferiority margin
  -> Inefficient: key toxicity outcomes are relegated to secondary analyses

⚠ Does not answer the clinical question: "which drug patients are better off"

⚠ Can produce results where the reduced dose is less toxic and the full dose does not lead to statistically superior survival, yet one does not have convincing evidence for non-inferiority.



Laken et al., AMPPS (2018)

Benefit-risk assessment using GPC

Prioritized outcomes:

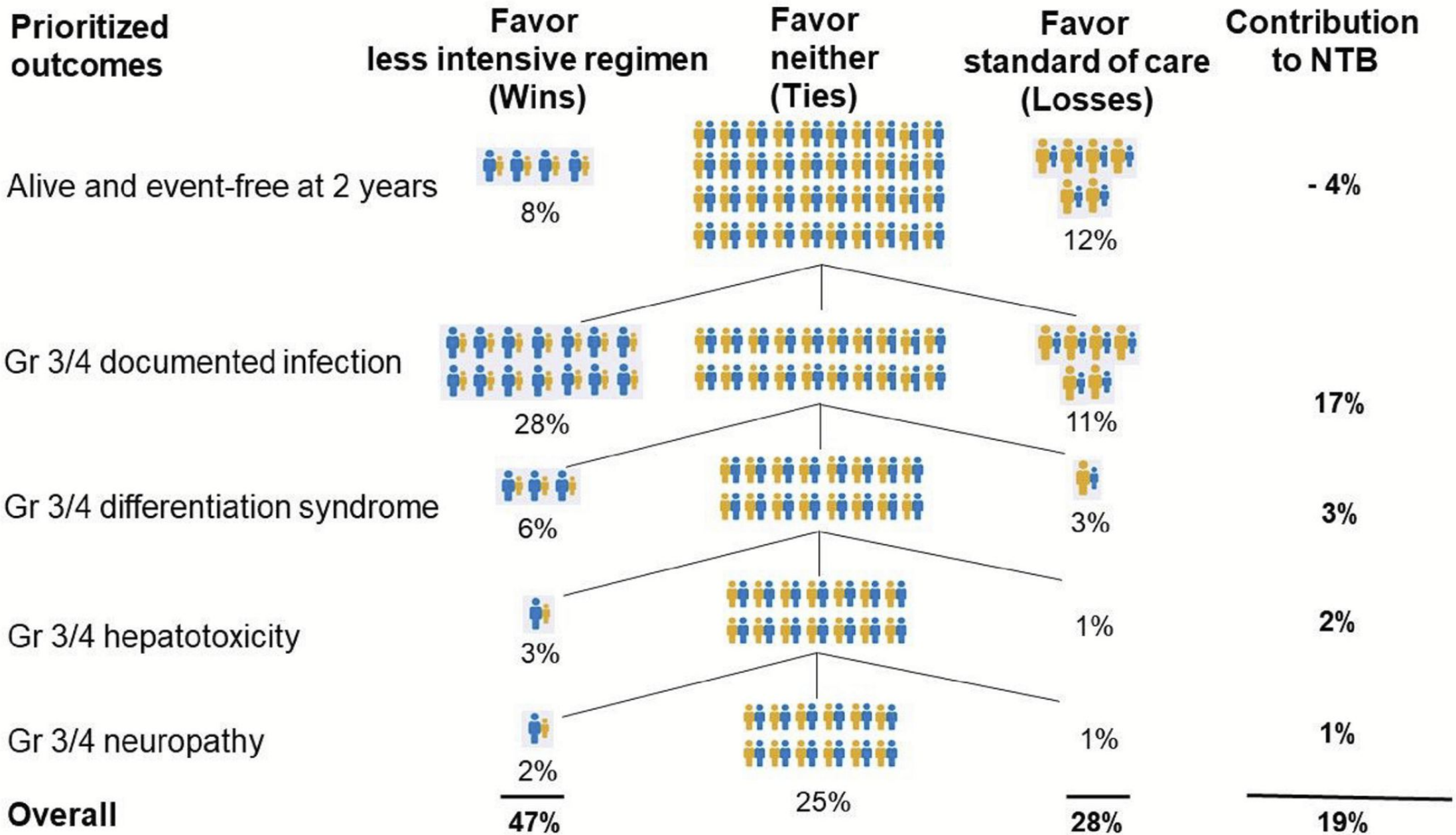
- Event free survival with the non-inferiority margin as threshold (e.g. 1 month)
- Severe side effects (e.g. grade 3 or 4)
- … any other outcome **relevant** for the patient or regulator. *Tannock et. al, Lancet Oncol 2024*

Toxicity is only investigated among pairs of patients with similar survival

- order of the outcomes reflects their importance
- accounts for potential degradation of the survival with the reduced dose

Testing Net Treatment Benefit for reduced vs. full dose:

- more meaningful: *Does benefit-risk balance favor reduced dose?*
- typically better powered



De Backer, Clin Trials (2024)

Challenges in designing trials with GPC

Definition of the estimand:

- ordering of the outcomes
- threshold(s) of clinical relevance
- encoding of the outcome: typically more neutral pairs when categorical vs. continuous
  more neutral pairs => more weight for later outcomes

interim analyses may lead to different estimands, e.g. with survival as first outcome
more neutral pairs at interim compared to final due to shorter follow-up time

Simulations studies are typically required

- GPC being non-parametric: no explicit formula except special cases
- joint distribution required: hypothesis on the outcome dependence structure

Power/sample size calculation with BuyseTest

Step 1: define the data-generating mechanism via a function

- generate a data.frame, one row per subject, one column for group and for each outcome
- simplistic example with independent outcomes

no censoring, no terminal event

R code

```
> simFCT <- function(n.C, n.T){
   df.C <- data.frame(id = paste0("C",1:n.C), group = 0,
                      tox = sample(1:6, n.C, replace=TRUE,
                                   prob = c(16.09, 15.42, 33.26, 26.18, 8.38, 0.67)/100),
                      time = rweibull(n.C, scale = 9.995655, shape = 1.28993),
                      event = 1)
   df.T <- data.frame(id = paste0("T",1:n.T), group = 1,
                      tox = sample(1:6, n.T, replace=TRUE,
                                   prob = c(8.21, 13.09, 31.29, 30.87, 12.05, 4.49)/100),
                      time = rweibull(n.T, scale = 13.16543, shape = 1.575269),
                      event = 1)
   return(rbind(df.C,df.T))
}
> set.seed(10)
> simFCT(2,2)
```

`data.frame` format

R output

	id	group	tox	time	event
1	C1	0	4	8.821945	1
2	C2	0	3	4.591318	1
3	T1	1	3	15.495787	1
4	T2	1	3	15.557655	1

Power calculation with BuyseTest

Step 2: use the function `powerBuyseTest` to evaluate rejection rate under the proposed data generating mechanism

R code

```
> e.power <- powerBuyseTest(group ~ tte(time,event,threshold = 1)  + cont(tox, operator = "<0"),
                            sim = simFCT, sample.size = c(10,25,50),
                            n.rep = 100, seed = 10, cpus = 1)
> summary(e.power)
```

R output

```
        Simulation study with Generalized pairwise comparison with 100 samples

- net benefit statistic (null hypothesis Delta=0)
endpoint threshold n.T n.C mean.estimate sd.estimate mean.se rejection.rate
     tox     1e-12  10  10        0.2156      0.2656  0.2468           0.13
                    25  25        0.2032      0.1677  0.1582            0.2
                    50  50        0.2015      0.1228  0.1121           0.43

n.T           : number of observations in the treatment group
n.C           : number of observations in the control group
mean.estimate: average estimate over simulations
sd.estimate   : standard deviation of the estimate over simulations
mean.se       : average estimated standard error of the estimate over simulations
rejection     : frequency of the rejection of the null hypothesis over simulations
(standard error: H-projection of order 1| p-value: after transformation)
```

number of simulations used to estimate the rejection rate

⚠ should be larger to get precise results

Sample size calculation with BuyseTest

Step 2: use the function `powerBuyseTest` to approximate the sample size

(based on an asymptotic approximation, not accurate with small sample size)

R code

```
> e.n <- powerBuyseTest(group ~ tte(time,event, threshold = 1) + cont(tox, operator = "<0"),
                        sim = simFCT, power = 0.8,
                        n.rep = c(1000,10), seed = 10, trace = 2, cpus = 1)
> summary(e.n)
```

R output

```
        Sample size calculation with Generalized pairwise comparison
        for a power of 0.8 and type 1 error rate of 0.05

- estimated sample size (mean [min;max]): 126 [91;155] controls
126 [91;155] treated

- net benefit statistic (null hypothesis Delta=0)
endpoint threshold n.T n.C mean.estimate sd.estimate mean.se rejection.rate
     tox     1e-12 126 126        0.2049       0.069  0.0707          0.818

n.T          : number of observations in the treatment group
n.C          : number of observations in the control group
mean.estimate: average estimate over simulations
sd.estimate  : standard deviation of the estimate over simulations
mean.se      : average estimated standard error of the estimate over simulations
rejection    : frequency of the rejection of the null hypothesis over simulations
(standard error: H-projection of order 1| p-value: after transformation)
```

10 large datasets (default n=m=2000) used to estimate the asymptotic variance

1000 simulations used to estimate the rejection rate

GPC in a nutshell

Principled way to combine outcomes of different nature

- hierarchy, threshold of clinical relevant, restriction time
- require careful considerations and discussion

to address a pertinent clinical question

- patient centric: what treatment benefits most the patient

while understanding the impact of each outcome

- Net Treatment Benefit as an interpretable and additive effect measure

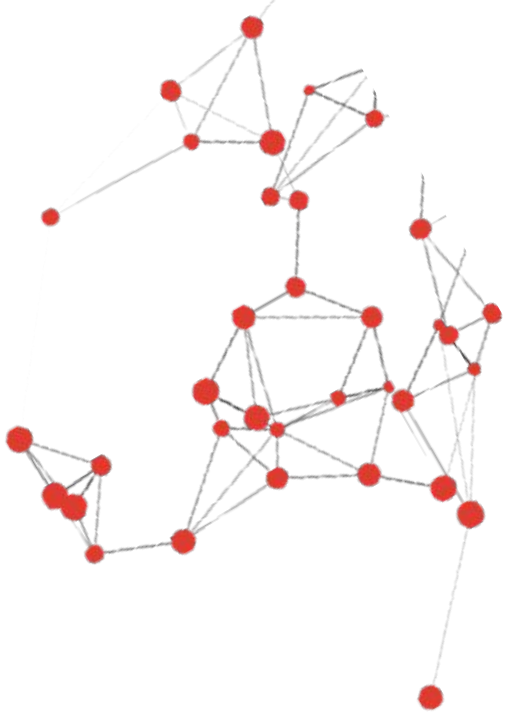
A ‘mature’ framework that can handle many of the usual complications

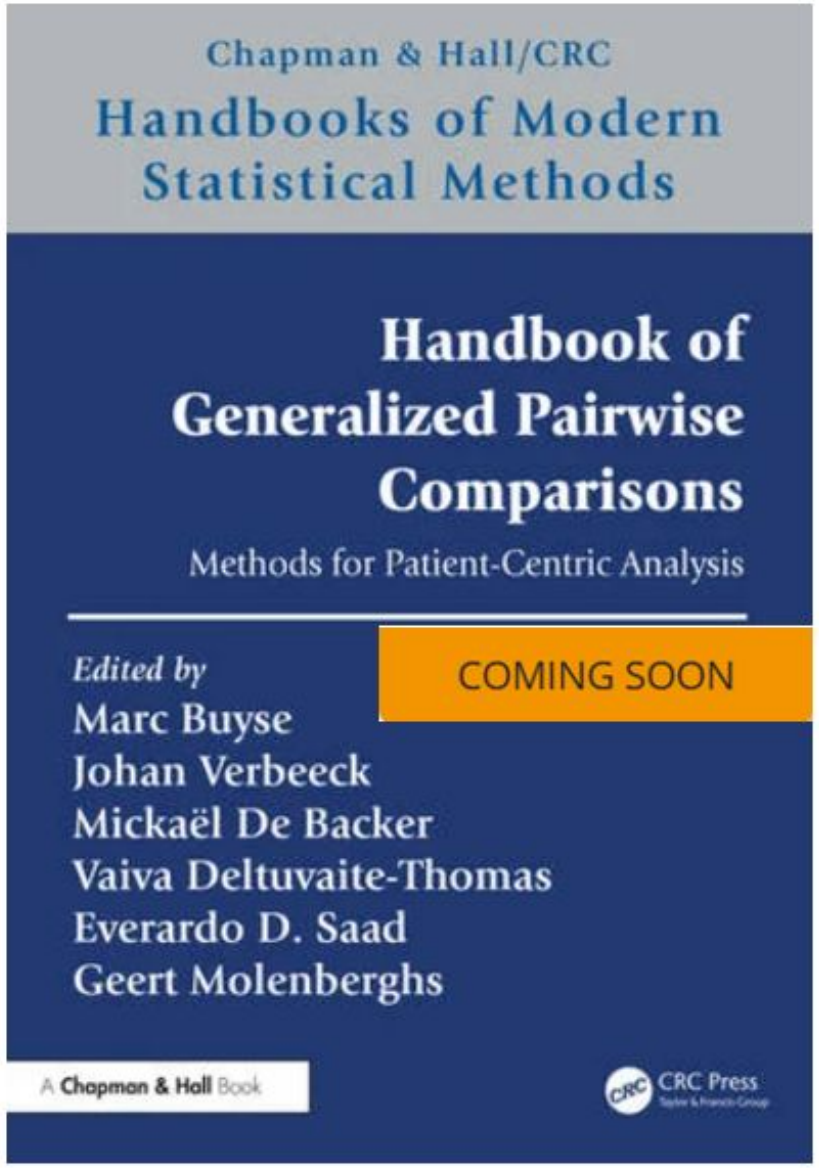
- right-censoring, competing risks, covariates, multiple testing
- The  package BuyseTest attempts to provide a convenient & transparent interface

despite remaining open questions

- interim analyses, non-transitivity with >2 groups, correlated right-censored outcomes, …

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

Bibliography (1/6)

Anderson,W.N.; Verbeeck, J. Exact Permutation and Bootstrap Distribution of Generalized Pairwise Comparisons Statistics. Mathematics 2023, 11, 1502. https://doi.org/10.3390/math11061502

Buyse M. Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. Stat Med. 2010 Dec 30;29(30):3245-57. doi: 10.1002/sim.3923.

Brunner E, Vandemeulebroecke M, and Mütze T. Win odds: An adaptation of the win ratio to include ties. Statistics in Medicine. 2021; 40, 3367–3384.

Brunner, E., Konietschke, F. An unbiased rank-based estimator of the Mann–Whitney variance including the case of ties. *Stat Papers* **66**, 20 (2025). https://doi.org/10.1007/s00362-024-01635-0

DATTA, S., BANDYOPADHYAY, D. and SATTEN, G.A. (2010), Inverse Probability of Censoring Weighted *U*-statistics for Right-Censored Data with an Application to Testing Hypotheses. Scandinavian Journal of Statistics, 37: 680-700. https://doi.org/10.1111/j.1467-9469.2010.00697.x

De Backer M, Legrand C, Péron J, Lambert A, Buyse M. On the use of extreme value tail modeling for generalized pairwise comparisons with censored outcomes. *Pharmaceutical Statistics*. 2023; 22(2): 284-299. doi:10.1002/pst.2271

De Backer M, Sengar M, Mathews V, et al. Design of a clinical trial using generalized pairwise comparisons to test a less intensive treatment regimen. *Clinical Trials*. 2024;21(2):180-188. doi:10.1177/17407745231206465

Deltuvaite-Thomas, V., Verbeeck, J., Burzykowski, T., Buyse, M., Tournigand, C., Molenberghs, G., & Thas, O. (2023). Generalized pairwise comparisons for censored data: An overview. *Biometrical Journal*, 65, 2100354. https://doi.org/10.1002/bimj.202100354

Bibliography (2/6)

Dong G, Hoaglin DC, Qiu J, Matsouaka RA, Chang YW, Wang J, Vandemeulebroecke M. The Win Ratio: On Interpretation and Handling of Ties. Statistics in Biopharmaceutical Research. 2019; 12(1), 99–106. https://doi.org/10.1080/19466315.2019.1575279

Dong Gaohong, Mao Lu , Huang Bo, Gamalo-Siebers Margaret , Wang Jiuzhou, Yu GuangLei & Hoaglin David C. (2020) The inverse-probability-of-censoring weighting (IPCW) adjusted win ratio statistic: an unbiased estimator in the presence of independent censoring, Journal of Biopharmaceutical Statistics, 30:5, 882-899, DOI:10.1080/10543406.2020.1757692

Dong G, Hoaglin DC, Huang B, Cui Y, Wang D, Cheng Y, Gamalo-Siebers M. The stratified win statistics (win ratio, win odds, and net benefit). Pharm Stat. 2023 Jul-Aug;22(4):748-756. doi: 10.1002/pst.2293.

Efron, Bradley, The two sample problem with censored data, Berkeley Symp. on Math. Statist. and Prob., 1967 5.4: 831-853 (1967)

Fay MP, Brittain EH, Shih JH, Follmann DA, Gabriel EE. Causal estimands and confidence intervals associated with Wilcoxon-Mann-Whitney tests in randomized experiments. Stat Med. 2018 Sep 10;37(20):2923-2937. doi: 10.1002/sim.7799.

Gehan, Edmund A. “A Generalized Two-Sample Wilcoxon Test for Doubly Censored Data.” *Biometrika* 52, no. 3/4 (1965): 650–53. https://doi.org/10.2307/2333721.

Hoeffding W. A Class of Statistics with Asymptotically Normal Distribution. Ann. Math. Statist. 1948 Sept; 19 (3): 293 - 325. https://doi.org/10.1214/aoms/1177730196

Jaspers S, Verbeeck J, Thas O. Covariate-adjusted generalized pairwise comparisons in small samples. Statistics in Medicine. 2024; 43(21): 4027-4042. doi: 10.1002/sim.10140

Bibliography (3/6)

Koch GG, Tangen CM, Jung JW, Amara IA. Issues for covariance analysis of dichotomous and ordered categorical data from randomized clinical trials and non-parametric strategies for addressing them. Stat Med. 1998 Aug 15-30;17(15-16):1863-92. doi: 10.1002/(sici)1097-0258(19980815/30)17:15/16<1863::aid-sim989>3.0.co;2-m.

Konietschke F, Hothorn LA, Brunner E. Rank-based multiple test procedures and simultaneous confidence intervals. Electron. J. Statist. 2012; 6: 738 - 759. https://doi.org/10.1214/12-EJS691

Lakens D, Scheel AM, Isager PM. Equivalence Testing for Psychological Research: A Tutorial. Advances in Methods and Practices in Psychological Science. 2018;1(2):259-269. doi:10.1177/2515245918770963

Mann HB, Whitney, DR. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. The Annals of Mathematical Statistics. 1947; 18(1), 50–60.

Mao L. On causal estimation using U-statistics. Biometrika. 2018; 105(1): 215–220, https://doi.org/10.1093/biomet/asx071

Mao L, Wang T. A class of proportional win-fractions regression models for composite outcomes. Biometrics. 2021; 77: 1265–1275. https://doi.org/10.1111/biom.13382

Matsouaka RA. Robust statistical inference for matched win statistics. Stat Methods Med Res. 2022 Aug;31(8):1423-1438. doi: 10.1177/09622802221090761.

Bibliography (4/6)

O'Brien PC. Procedures for comparing samples with multiple endpoints. Biometrics. 1984 Dec;40(4):1079-87.

Ozenne B, Budtz-Jørgensen E, Péron J. The asymptotic distribution of the Net Benefit estimator in presence of right-censoring. Stat Methods Med Res. 2021 Nov;30(11):2399-2412. doi: 10.1177/09622802211037067

Péron J, Buyse M, Ozenne B, Roche L, Roy P. An extension of generalized pairwise comparisons for prioritized outcomes in the presence of censoring. Stat Methods Med Res. 2018 Apr;27(4):1230-1239. doi: 10.1177/0962280216658320

Péron J, Giai J, Maucort-Boulch D, Buyse M. The Benefit-Risk Balance of Nab-Paclitaxel in Metastatic Pancreatic Adenocarcinoma. Pancreas. 2019 Feb;48(2):275-280. doi: 10.1097/MPA.0000000000001234

Péron J, Idlhaj M, Maucort-Boulch D, Giai J, Roy P, Collette L, Buyse M, Ozenne B. Correcting the bias of the net benefit estimator due to right-censored observations. Biom J. 2021 Apr;63(4):893-906. doi: 10.1002/bimj.202000001

Piffoux M, Ozenne B, De Backer M, Buyse M, Chiem JC, Péron J. Restricted Net Treatment Benefit in oncology. J Clin Epidemiol. 2024 Jun;170:111340. doi: 10.1016/j.jclinepi.2024.111340

Pocock SJ, Ariti CA, Collier TJ, Wang D. The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. Eur Heart J. 2012 Jan;33(2):176-82. doi: 10.1093/eurheartj/ehr352.

Rosner BA. Fundamentals of Biostatistics. Boston :Brooks/Cole, Cengage Learning, 1999.

Bibliography (5/6)

Tannock IF, Buyse M, De Backer M, Earl H, Goldstein DA, Ratain MJ, Saltz LB, Sonke GS, Strohbehn GW. The tyranny of non-inferiority trials. Lancet Oncol. 2024 Oct;25(10):e520-e525. doi: 10.1016/S1470-2045(24)00218-3

Thas O, Neve JD, Clement L, Ottoy JP. Probabilistic index models. Journal of the Royal Statistical Society: Series B.2012; 74: 623-671. https://doi.org/10.1111/j.1467-9868.2011.01020.x

Verbeeck J, De Backer M, Verwerft J, Salvaggio S, Valgimigli M, Vranckx P, Buyse M, Brunner E. Generalized Pairwise Comparisons to Assess Treatment Effects: JACC Review Topic of the Week. J Am Coll Cardiol. 2023 Sep 26;82(13):1360-1372. doi: 10.1016/j.jacc.2023.06.047.

Verbeeck J, Dirani M, Bauer JW, Hilgers RD, Molenberghs G, Nabbout R. Composite endpoints, including patient reported outcomes, in rare diseases. Orphanet J Rare Dis. 2023 Sep 1;18(1):262. doi: 10.1186/s13023-023-02819-x.

Verbeeck J, Saad ED. Rethinking survival analysis: advancing beyond the hazard ratio? Eur Heart J Acute Cardiovasc Care. 2024 Mar 11;13(3):313-315. doi: 10.1093/ehjacc/zuae017.

Vermeulen K, Thas O, Vansteelandt S. Increasing the power of the Mann-Whitney test in randomized experiments through flexible covariate adjustment. Statist. Med. 2015; 34: 1012–1030. doi: 10.1002/sim.6386

Vermeulen K, Vansteelandt S. Data-Adaptive Bias-Reduced Doubly Robust Estimation. The International Journal of Biostatistics. 2016; 12(1): 253-282. https://doi.org/10.1515/ijb-2015-0029

Bibliography (6/6)

Von Hoff DD, Ervin T, Arena FP, Chiorean EG, Infante J, Moore M, Seay T, Tjulandin SA, Ma WW, Saleh MN, Harris M, Reni M, Dowden S, Laheru D, Bahary N, Ramanathan RK, Tabernero J, Hidalgo M, Goldstein D, Van Cutsem E, Wei X, Iglesias J, Renschler MF. Increased survival in pancreatic cancer with nab-paclitaxel plus gemcitabine. N Engl J Med. 2013 Oct 31;369(18):1691-703. doi: 10.1056/NEJMoa1304369

Wally V, Hovnanian A, Ly J, Buckova H, Brunner V, Lettner T, Ablinger M, Felder TK, Hofbauer P, Wolkersdorfer M, Lagler FB, Hitzl W, Laimer M, Kitzmüller S, Diem A, Bauer JW. Diacerein orphan drug development for epidermolysis bullosa simplex: A phase 2/3 randomized, placebo-controlled, double-blind clinical trial. J Am Acad Dermatol. 2018 May;78(5):892-901.e7. doi: 10.1016/j.jaad.2018.01.019.

Wilcoxon F. Individual Comparisons by Ranking Methods. Biometrics Bulletin. 1945; 1(6): 80–83. https://doi.org/10.2307/3001968

Yusuf S, Pfeffer MA, Swedberg K, Granger CB, Held P, McMurray JJ, Michelson EL, Olofsson B, Ostergren J; CHARM Investigators and Committees. Effects of candesartan in patients with chronic heart failure and preserved left-ventricular ejection fraction: the CHARM-Preserved Trial. Lancet. 2003 Sep 6;362(9386):777-81. doi: 10.1016/S0140-6736(03)14285-7.

Zhang Z, Ma S, Shen C, Liu C. Estimating Mann–Whitney-type Causal Effects. International Statistical Review. 2019; 87: 514–530. https://doi.org/10.1111/insr.12326.