Using Generalized Pairwise Comparisons to compare the benefit-risk profile of two therapies

Brice Ozenne[1,2]
collaboration with Johan Verbeeck, Marc Buyse, Julien Péron

[1] Neurobiology Research Unit, University Hospital of Copenhagen, Rigshospitalet.

[2] Section of Biostatistics, Department of Public Health, University of Copenhagen.

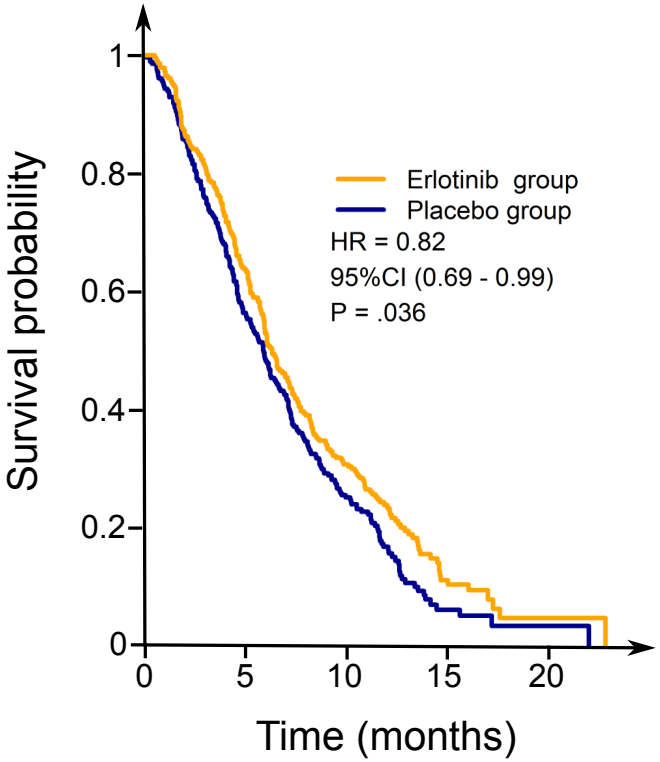
Online seminar, 29-04-2026
Institute of Biostatistics and Clinical Research of the University of Münster

Motivations

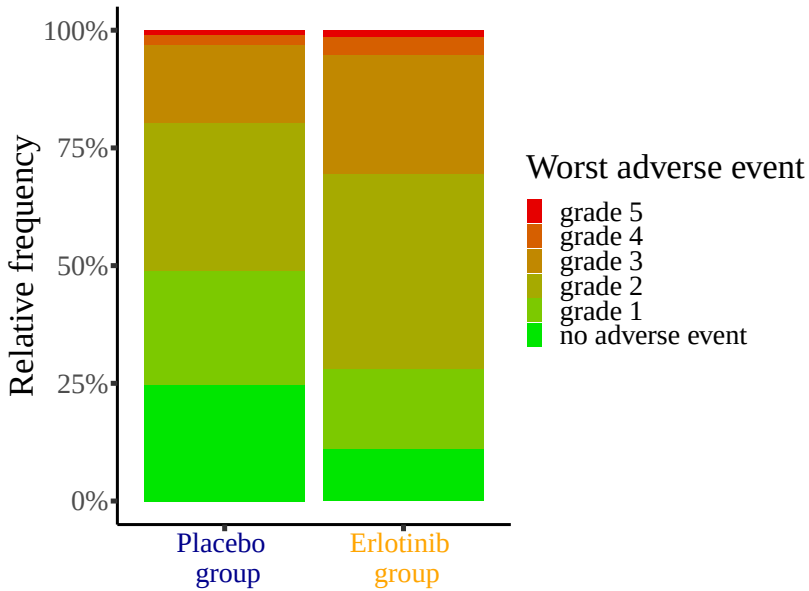
- Benefit-risk quantification
  - A Wilcoxon-Mann-Whitney test ++

Example of clinical trial: adapted from Moore et al. (2007)

Benefit (time to death)



Risk (non-lethal adverse event)



- survival benefit (statistically significant)
- more serious adverse events

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

Example of clinical trial: from Von Hoff et al. (2013)

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.,

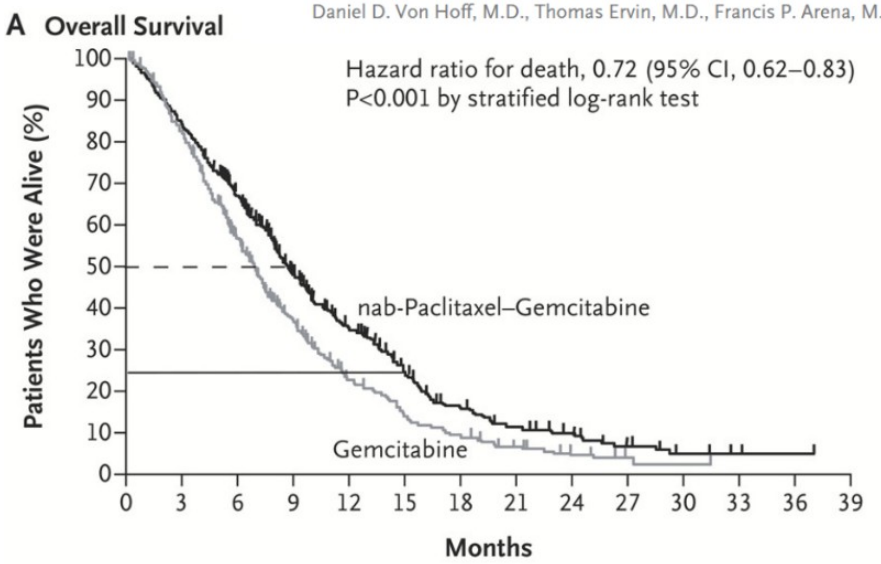


Table 3. Common Adverse Events of Grade 3 or Higher and Growth-Factor Use.*

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

- survival benefit (statistically significant)
- more serious adverse events 😒

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

Limitation of the traditional approach

Marginal benefit risk analyses:

- Benefit: ~~log-rank test~~
  difference in 1 year survival.
- Risk: ~~chi-squared test~~
  difference in proportion of patients with serious side effects

Limitation of the traditional approach

Marginal benefit risk analyses:

- Benefit: ~~log-rank test~~
  difference in 1 year survival.
- Risk: ~~chi-squared test~~
  difference in proportion of patients with serious side effects

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.

Illustration of scenario a), b), and c)

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

Illustration of scenario a), b), and c)

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 should provide valuable information $\rightarrow$ GPC.

What about time-to-first event analyses?

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 Michelson, 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%)

Events in time-to-first event composite

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

What about time-to-first event analyses?

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 Michelson, 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%)

Events in time-to-first event composite

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

Ignores 46% (158/340) of CV deaths

⚠ Emphasis is on time of event, rather than severity of event

- a patient that has an hospitalization is worse than a patient dying 1 day later

Extending the Wilcoxon-Mann-Whitney test

Generalized Pairwise Comparisons (GPC) can be seen as an extension of the Wilcoxon-Mann-Whitney test.

GPC is able to handle:

- multiple outcomes (e.g. survival, toxicity)
- covariates (e.g. stratify on gender)
- missing data (e.g. right-censoring)
- competing risks (e.g. death)
- heteroschedasticity (e.g. more variable treatment group)

 see "Handbook of GPC" by Buyse et al. (2025) for more details.

Wilcoxon-Mann-Whitney test with an estimand

GPC makes the analysis more transparent

- estimate(s) that can be decomposed per outcome

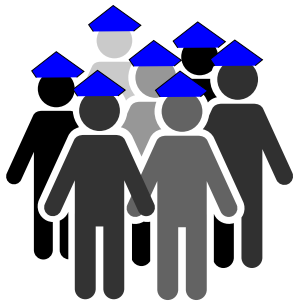
Prioritized Outcomes	Favorable values	Pairs favoring Experiment	Neutral pairs	Pairs favoring Control	Contribution to NTB	NTB (*p*-value)	Information proportion (cumulative)
Grade 4 SOM	No (vs. Yes)	25.7%	58.0%	16.3%	9.4%	9.4% ($p = 0.040$)	42.0%
Grade 3 SOM	No (vs. Yes)	14.0%	33.3%	10.7%	3.3%	12.7% ($p = 0.020$)	24.7% (66.7%)
SOM duration	Shorter (by ≥ 1 week)	10.0%	17.2%	6.2%	3.8%	16.5% ($p = 0.0033$)	16.2% (82.9%)
Time to SOM	Later (by ≥ 1 week)	4.2%	11.0%	1.9%	2.3%	18.8% ($p = 0.0012$)	6.1% (89.0%)
Overall		53.9%		35.1%		**18.8% ($p = 0.0012$)**	

- causal interpretation can be made clear (Fay et al., 2018)

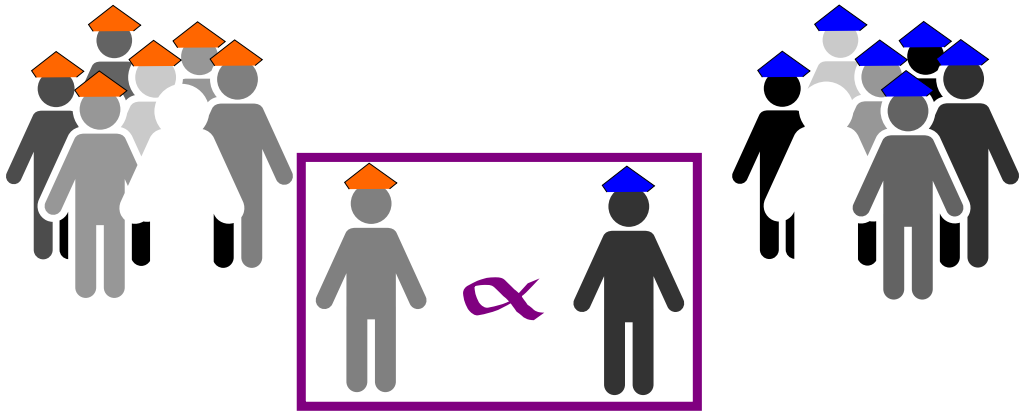
Generalized Pairwise Comparisons (GPC)

- estimator
- 4 summary statistics, 1 recommended.
- relation to traditional tests/effect size measures
- causal interpration & limitations
- implementation with the  package BuyseTest

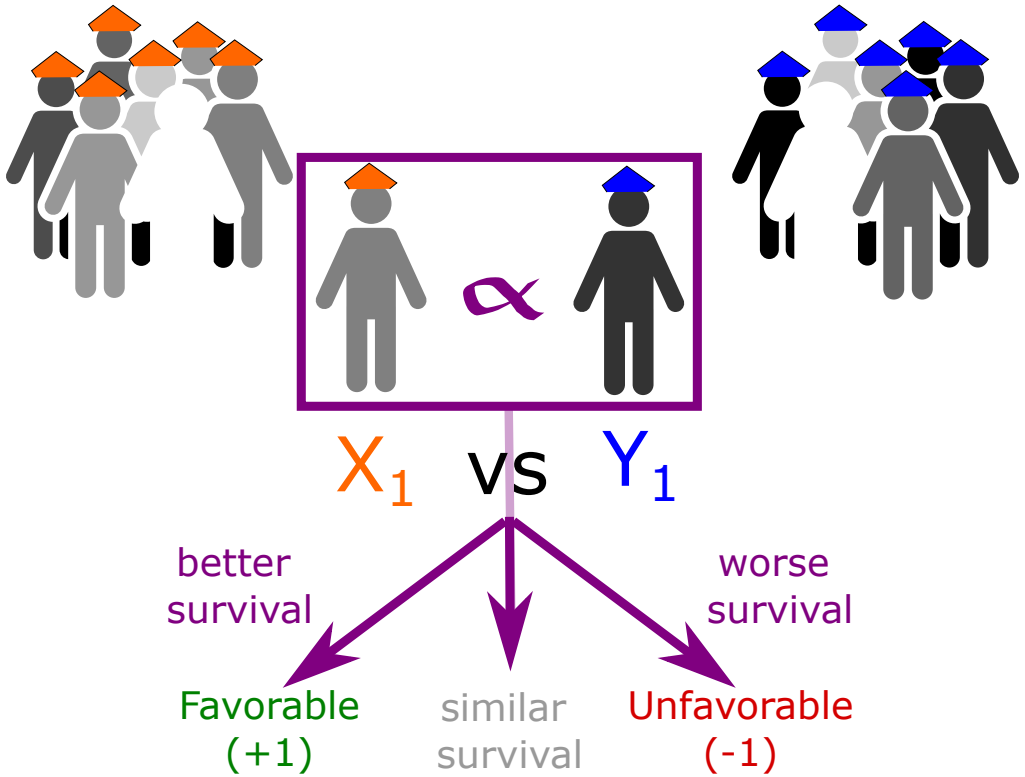
Illustrating GPC



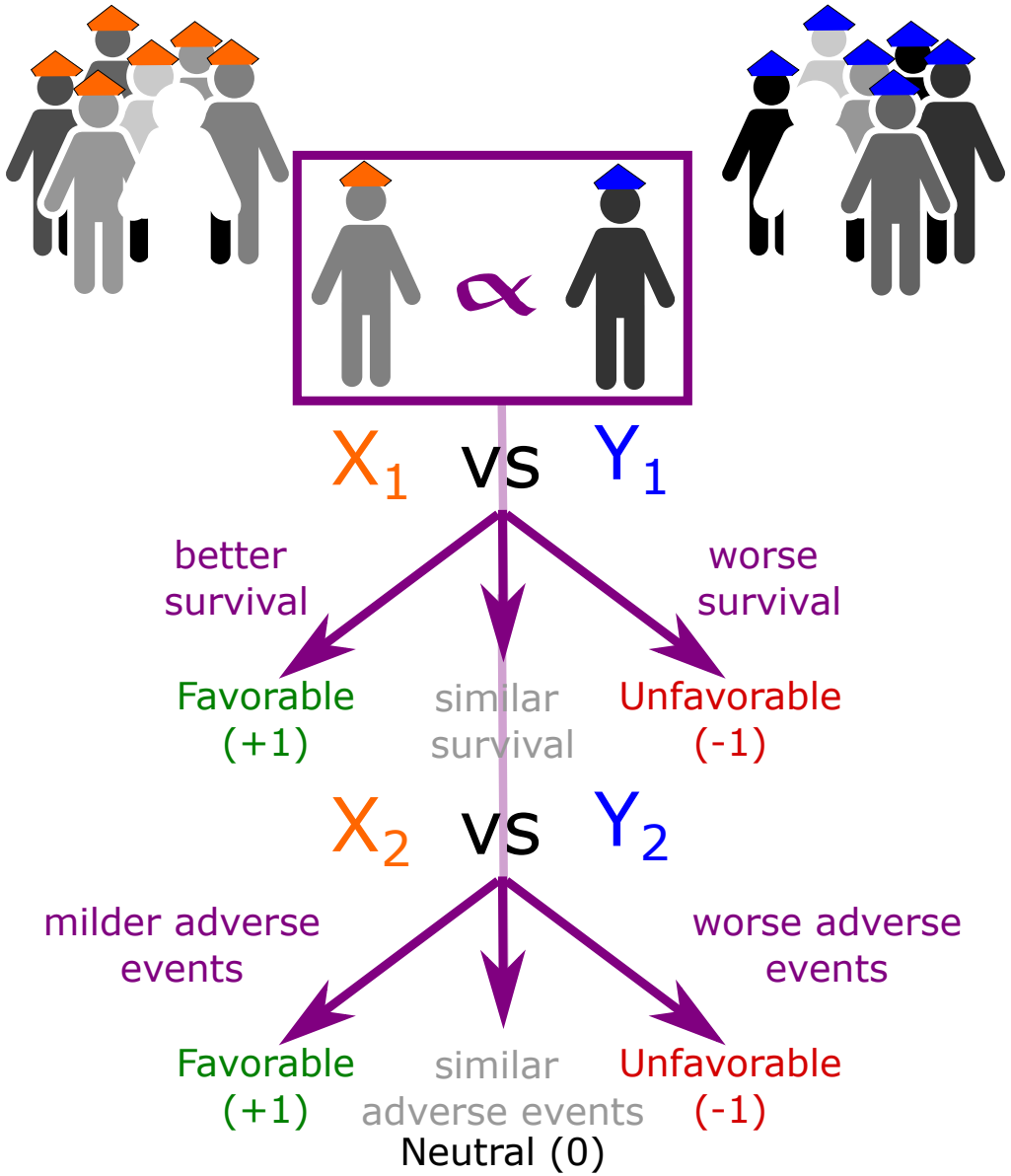
Illustrating GPC



Illustrating GPC



Illustrating GPC



GPC estimator in a nutshell

1. Decide on a hierarchy of outcomes and associated threshold of clincal relevance, e.g.:
   - first priority: outcome (X_1, Y_1) with threshold τ_1
   - second priority: outcome (X_2, Y_2) with threshold τ_2
2. Form all distinct pairs with one subject from each group (i,j) and score each pair, e.g.:

$$
\begin{aligned}
U_{i,j}^{+} &= \mathbb{1}_{X_{1i} \geq Y_{1j}+\tau_1} + \mathbb{1}_{X_{2i} \geq Y_{2j}+\tau_2} \mathbb{1}_{|X_{1i}-Y_{1j}|<\tau_1} \\
U_{i,j}^{-} &= \mathbb{1}_{Y_{1j} \geq X_{1i}+\tau_1} + \mathbb{1}_{Y_{2j} \geq X_{2i}+\tau_2} \mathbb{1}_{|X_{1i}-Y_{1j}|<\tau_1} \\
U_{i,j}^{0} &= \mathbb{1}_{|X_{2i}-Y_{2j}|<\tau_2} \mathbb{1}_{|X_{1i}-Y_{1j}|<\tau_1}
\end{aligned}
$$

3. Summarize the scores across subjects into a single statistic:

$$
\widehat{\Delta} = \frac{1}{n_X n_Y} \sum_{i=1}^{n_X} \sum_{j=1}^{n_Y} U_{i,j}^{+} - U_{i,j}^{-}
$$

package BuyseTest

```
library(BuyseTest)
data(prodige)

e.BT <- BuyseTest(
  treatment ~ tte(OS, statusOS, threshold = 5)
              + cont(toxicity, operator = "<0"),
  data = prodige
)
print(model.tables(e.BT)[,c(1:6,9)], digits = 3)
```

```
  endpoint threshold total favorable unfavorable neutral Delta
1       OS     5e+00 100.0      30.8        18.0    51.1 0.127
3 toxicity     1e-12  51.2      17.4        19.2    14.6 0.110
```

Equivalence with Wilcoxon-Mann-Whitney test

Single outcome and infinitesimal threshold τ of clinical relevance

```
## simulate data
set.seed(1)
df <- data.frame(Y = rnorm(100), E = rep(c("A","B"),50))
```

```
## Wilcoxon test
wilcox.test(Y ~ E, data = df, correct = FALSE)$p.value
```

```
[1] 0.3276183
```

```
## Net Treatment Benefit test via GPC
resB <- BuyseTest(E ~ cont(Y) , data = df,
     method.inference = "varexact-permutation")
confint(resB)[,c("estimate","se","p.value")]
```

```
  estimate       se   p.value
Y  -0.1136 0.116046 0.3276183
```

GPC summary statistic

Averaging the scores over all pairs, we can consider $U^{+}_{\bullet,\bullet}$, $U^{-}_{\bullet,\bullet}$, $U^{0}_{\bullet,\bullet}$ as sufficient statistics.

In step 3, we considered the **Net Treatment Benefit (NTB)** as summary statistic

- take the difference between $U^{+}_{\bullet,\bullet}$ and $U^{-}_{\bullet,\bullet}$ and ignores $U^{0}_{\bullet,\bullet}$.
- analogue to a correlation:

$$\widehat{\Delta} = \begin{cases} 1, \text{ treatment always better} \\ 0, \text{ no difference in average} \\ -1, \text{ treatment always worse} \end{cases}$$

GPC summary statistic

Averaging the scores over all pairs, we can consider $U^{+}_{\bullet,\bullet}$, $U^{-}_{\bullet,\bullet}$, $U^{0}_{\bullet,\bullet}$ as sufficient statistics.

In step 3, we considered the **Net Treatment Benefit (NTB)** as summary statistic

- take the difference between $U^{+}_{\bullet,\bullet}$ and $U^{-}_{\bullet,\bullet}$ and ignores $U^{0}_{\bullet,\bullet}$.
- analogue to a correlation:

$$\widehat{\Delta} = \begin{cases} 1, \text{ treatment always better} \\ 0, \text{ no difference in average} \\ -1, \text{ treatment always worse} \end{cases}$$

Other summary statistics have been proposed in the litterature:

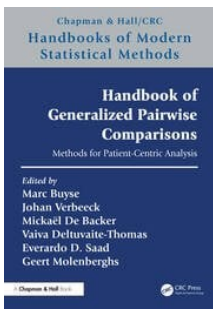
- **Probabilistic index (PI)**: $U^{+}_{\bullet,\bullet} + 0.5U^{0}_{\bullet,\bullet}$
- **Win Ratio (WR)** $\frac{U^{+}_{\bullet,\bullet}}{U^{-}_{\bullet,\bullet}}$
- **Success Odds (SO)** $\frac{U^{+}_{\bullet,\bullet}+0.5U^{0}_{\bullet,\bullet}}{U^{-}_{\bullet,\bullet}+0.5U^{0}_{\bullet,\bullet}}$

GPC summary statistic - relationships

Relationships between the GPC-based treatment effects, Probabilistic Index (PI), Net Treatment Benefit (NTB), Success Odds (SO) and Win Ratio (WR). The table is written such that each effect reported in a line is written as a function of another effect reported in a column, provided there is indeed an analytical relationship between the two.

	PI	NTB	SO	WR
PI	Θ_τ	$\Theta_\tau = \frac{\Delta_\tau + 1}{2}$	$\Theta_\tau = \frac{\Lambda_\tau}{\Lambda_\tau + 1}$	-
NTB	$\Delta_\tau = 2\Theta_\tau - 1$	Δ_τ	$\Delta_\tau = \frac{\Lambda_\tau - 1}{\Lambda_\tau + 1}$	-
SO	$\Lambda_\tau = \frac{\Theta_\tau}{1 - \Theta_\tau}$	$\Lambda_\tau = \frac{1 + \Delta_\tau}{1 - \Delta_\tau}$	Λ_τ	-*
WR	-	-	-*	Ψ_τ

From



chapter 2

*: for settings where $\mathbb{P}(\boldsymbol{Y}^E \quad_\tau \boldsymbol{Y}^C) = 0$, Ψ_τ and Λ_τ are equivalent (e.g., for continuous distributions with no consideration of τ).

Shortcoming of the Win Ratio

Artificial example:

	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 (implicitely) redistributes the ties according to the observed win/loss proportions:

- overestimation of the effect

Additive decomposition of the NTB

GPC scores can be decomposed by outcome:

$$
\begin{aligned}
U^{+}_{i,j} &= \mathbb{1}_{X_{1i} \geq Y_{1j} + \tau_1} + \mathbb{1}_{X_{2i} \geq Y_{2j} + \tau_2} \mathbb{1}_{|X_{1i} - Y_{1j}| < \tau_1} + \dots \\
&= U^{+}_{i,j,1} + U^{+}_{i,j,2} + \dots \\
U^{-}_{i,j} &= \mathbb{1}_{Y_{1j} \geq X_{1i} + \tau_1} + \mathbb{1}_{Y_{2j} \geq X_{2i} + \tau_2} \mathbb{1}_{|X_{1i} - Y_{1j}| < \tau_1} + \dots \\
&= U^{-}_{i,j,1} + U^{-}_{i,j,2} + \dots
\end{aligned}
$$

Additive decomposition of the NTB

GPC scores can be decomposed by outcome:

$$\begin{aligned} U_{i,j}^{+} &= \mathbb{1}_{X_{1i} \geq Y_{1j} + \tau_1} + \mathbb{1}_{X_{2i} \geq Y_{2j} + \tau_2} \mathbb{1}_{|X_{1i} - Y_{1j}| < \tau_1} + \dots \\ &= U_{i,j,1}^{+} + U_{i,j,2}^{+} + \dots \\ U_{i,j}^{-} &= \mathbb{1}_{Y_{1j} \geq X_{1i} + \tau_1} + \mathbb{1}_{Y_{2j} \geq X_{2i} + \tau_2} \mathbb{1}_{|X_{1i} - Y_{1j}| < \tau_1} + \dots \\ &= U_{i,j,1}^{-} + U_{i,j,2}^{-} + \dots \end{aligned}$$

The same is true with the NTB (but not for the PI, WR, SO)

$$\begin{aligned} \widehat{\Delta} &= \frac{1}{n_X n_Y} \sum_{i=1}^{n_X} \sum_{j=1}^{n_Y} U_{i,j,1}^{+} - U_{i,j,1}^{-} + \frac{1}{n_X n_Y} \sum_{i=1}^{n_X} \sum_{j=1}^{n_Y} U_{i,j,2}^{+} - U_{i,j,2}^{-} + \dots \\ &= \widehat{\delta}_1 + \widehat{\delta}_2 + \dots \end{aligned}$$

Back to an example

Prioritized Outcomes	Favorable values	Pairs favoring Experiment	Neutral pairs	Pairs favoring Control	Contribution to NTB	NTB (*p*-value)	Information proportion (cumulative)
Grade 4 SOM	No (vs. Yes)	25.7%	58.0%	16.3%	9.4%	9.4% ($p = 0.040$)	42.0%
Grade 3 SOM	No (vs. Yes)	14.0%	33.3%	10.7%	3.3%	12.7% ($p = 0.020$)	24.7% (66.7%)
SOM duration	Shorter (by ≥ 1 week)	10.0%	17.2%	6.2%	3.8%	16.5% ($p = 0.0033$)	16.2% (82.9%)
Time to SOM	Later (by ≥ 1 week)	4.2%	11.0%	1.9%	2.3%	18.8% ($p = 0.0012$)	6.1% (89.0%)
Overall		53.9%		35.1%		**18.8% ($p = 0.0012$)**	

$$\text{NTB} = U^{+}_{\bullet,\bullet} - U^{-}_{\bullet,\bullet} \qquad \text{Information} = U^{+}_{\bullet,\bullet} + U^{-}_{\bullet,\bullet}$$

Back to an example

Prioritized Outcomes	Favorable values	Pairs favoring Experiment	Neutral pairs	Pairs favoring Control	Contribution to NTB	NTB (*p*-value)	Information proportion (cumulative)
Grade 4 SOM	No (vs. Yes)	25.7%	58.0%	16.3%	9.4%	9.4% ($p = 0.040$)	42.0%
Grade 3 SOM	No (vs. Yes)	14.0%	33.3%	10.7%	3.3%	12.7% ($p = 0.020$)	24.7% (66.7%)
SOM duration	Shorter (by ≥ 1 week)	10.0%	17.2%	6.2%	3.8%	16.5% ($p = 0.0033$)	16.2% (82.9%)
Time to SOM	Later (by ≥ 1 week)	4.2%	11.0%	1.9%	2.3%	18.8% ($p = 0.0012$)	6.1% (89.0%)
Overall		53.9%		35.1%		**18.8% ($p = 0.0012$)**	

$$\text{NTB} = U^{+}_{\bullet,\bullet} - U^{-}_{\bullet,\bullet} \qquad \text{Information} = U^{+}_{\bullet,\bullet} + U^{-}_{\bullet,\bullet}$$

⚠ The 'contribution to NTB' ($\widehat{\delta}_2 = 3.3\%$, $\widehat{\delta}_3 = 3.8\%$, $\widehat{\delta}_4 = 2.3\%$) should not be interpreted as a marginal effect.

- indeed $\widehat{\delta}_2$ compares the second outcome in the sub-population with similar first outcome values.

package BuyseTest

Default summary statistic: Net Treatment Benefit

```
print(model.tables(e.BT)[,c(1,3:6,8:9)], digits = 3)
```

```
  endpoint total favorable unfavorable neutral   delta Delta
1       OS 100.0      30.8        18.0    51.1  0.1275 0.127
3 toxicity  51.2      17.4        19.2    14.6 -0.0174 0.110
```

But you can choose your own, e.g.:

```
mytable <- model.tables(e.BT, statistic = "winRatio")
print(mytable[,c(1,3:6,8:9)], digits = 3)
```

```
  endpoint total favorable unfavorable neutral delta Delta
1       OS 100.0      30.8        18.0    51.1 1.707  1.71
3 toxicity  51.2      17.4        19.2    14.6 0.909  1.30
```

⚠ PI and SO requires adding the neutral score when calling BuyseTest (argument add.halfNeutral).

Underlying estimand

The GPC estimator of the **Net Treatment Benefit (NTB)**:

$$\widehat{\Delta} = \frac{1}{n_X n_Y} \sum_{i=1}^{n_X} \sum_{j=1}^{n_Y} U_{i,j}^{+} - U_{i,j}^{-}$$

is a two sample U-statistic estimator of degree $\{1, 1\}$ converging towards:

$$\begin{aligned}\Delta &= \mathbb{P}\left[X_1 \geq Y_1 + \tau_1\right] + \mathbb{P}\left[X_2 \geq Y_2 + \tau_2, |X_1 - Y_1| < \tau_1\right] \\ &- \left(\mathbb{P}\left[Y_1 \geq X_1 + \tau_1\right] + \mathbb{P}\left[Y_2 \geq X_2 + \tau_2, |X_1 - Y_1| < \tau_1\right]\right) \\ &= \mathbb{P}\left[\boldsymbol{X} \succ_\tau \boldsymbol{Y}\right] - \mathbb{P}\left[\boldsymbol{Y} \succ_\tau \boldsymbol{X}\right]\end{aligned}$$

The NTB is the *net* probability of a better outcome in the treatment group vs. the placebo group.

Underlying estimand - not the ideal one

The NTB is the *net* probability of a better outcome in the treatment group vs. the placebo group.

- probability that a random patient i from the treatment group has a better outcome than a random patient j from the placebo group, minus the probability of the opposite situation.

$$\Delta = \mathbb{P}\left[\boldsymbol{X}_i \succ_\tau \boldsymbol{Y}_j\right] - \mathbb{P}\left[\boldsymbol{Y}_j \succ_\tau \boldsymbol{X}_i\right]$$

Underlying estimand - not the ideal one

The NTB is the *net* probability of a better outcome in the treatment group vs. the placebo group.

- probability that a random patient i from the treatment group has a better outcome than a random patient j from the placebo group, minus the probability of the opposite situation.

$$\Delta = \mathbb{P}\left[\boldsymbol{X}_i \succ_\tau \boldsymbol{Y}_j\right] - \mathbb{P}\left[\boldsymbol{Y}_j \succ_\tau \boldsymbol{X}_i\right]$$

⚠ It is not the difference between the probability for a patient $\boldsymbol{k}$ to have a better outcome under treatment vs. under placebo .

$$\Delta \neq \mathbb{P}\left[\boldsymbol{X}_k \succ_\tau \boldsymbol{Y}_k\right] - \mathbb{P}\left[\boldsymbol{Y}_k \succ_\tau \boldsymbol{X}_k\right] = \phi$$

Underlying estimand - not the ideal one

The NTB is the *net* probability of a better outcome in the treatment group vs. the placebo group.

- probability that a random patient i from the treatment group has a better outcome than a random patient j from the placebo group, minus the probability of the opposite situation.

$$\Delta = \mathbb{P}\left[\boldsymbol{X}_i \succ_\tau \boldsymbol{Y}_j\right] - \mathbb{P}\left[\boldsymbol{Y}_j \succ_\tau \boldsymbol{X}_i\right]$$

⚠ It is not the difference between the probability for a patient $\boldsymbol{k}$ to have a better outcome under treatment vs. under placebo .

$$\Delta \neq \mathbb{P}\left[\boldsymbol{X}_k \succ_\tau \boldsymbol{Y}_k\right] - \mathbb{P}\left[\boldsymbol{Y}_k \succ_\tau \boldsymbol{X}_k\right] = \phi$$

ϕ is a causal estimand (contrast between counterfactuals from the same subject). It is (generally) not identifiable but can be bounded (Fay et al., 2018).

Underlying causal estimand

With a absolutely continuous outcome (no ties), Fay et al. (2018) showed that the causal estimand behind probabilistic index (PI) is:

$$\mathbb{P}\left[\boldsymbol{X}_i > \boldsymbol{Y}_j\right] = \mathbb{E}\left[\overline{F}(X_k) - \overline{F}(Y_k)\right] + 0.5$$

where $\overline{F}(.) = 0.5F_X(.) + 0.5F_Y(.)$ is the 'average' outcome distribution pooling the two potential outcomes distribution with equal weights.

- outcome scale: $X_k - Y_k$ shift in outcome value
- quantile scale: $\overline{F}(X_k) - \overline{F}(Y_k)$ shift in position in the population after taking treatment

Underlying causal estimand

With a absolutely continuous outcome (no ties), Fay et al. (2018) showed that the causal estimand behind probabilistic index (PI) is:

$$\mathbb{P}\left[\boldsymbol{X}_i > \boldsymbol{Y}_j\right] = \mathbb{E}\left[\overline{F}(X_k) - \overline{F}(Y_k)\right] + 0.5$$

where $\overline{F}(.) = 0.5F_X(.) + 0.5F_Y(.)$ is the 'average' outcome distribution pooling the two potential outcomes distribution with equal weights.

- outcome scale: $X_k - Y_k$ shift in outcome value
- quantile scale: $\overline{F}(X_k) - \overline{F}(Y_k)$ shift in position in the population after taking treatment

The Net Treatment Benefit is about a robust quantification of the expected improvement, not about whether a majority benefits from the treatment (e.g. see Hand's paradox in Fay et al. (2018)).

Relation to effect size measures

Consider the Net Treatment Benefit (NTB) with a single outcome and an infinitesimal threshold τ

- **Binary outcome** $\Delta = \mathbb{P}[X = 1] - \mathbb{P}[Y = 1]$
- **Time to event outcome**: $\Delta = \frac{1-HR}{1+HR}$ under proportional hazard and no censoring with HR being the hazard ratio.
- **Continuous outcome**: $\Delta = 2\Phi\left(\frac{d}{\sqrt{2}}\right) - 1$ under normally distributed outcome with d being Cohen's d and Φ the cumulative distribution function of a normal distribution.

Criticism of GPC

Summary statistics, like the Net Treatment Benefit (NTB), are trial specific

- e.g. with normally distributed outcomes the NTB depends on the outcome variance, i.e., on the inclusion criteria.

Stratum-specific analyses may disagree with marginal analyses:

- e.g. consider female and male strata: (♀, ♂) vs. (♀, ♂). Equal number of male and females.
  $\Delta(\bullet, \bullet) = \mathbb{P}[\bullet \succ_\tau \bullet] - \mathbb{P}[\bullet \succ_\tau \bullet]$

$$\underbrace{\frac{\Delta(♀, ♀) + \Delta(♂, ♂)}{2}}_{\text{Pooled stratified NTB}} \neq \underbrace{\frac{\Delta(♂, ♂) + \Delta(♀, ♀) + \Delta(♂, ♀) + \Delta(♀, ♂)}{4}}_{\text{marginal NTB}}$$

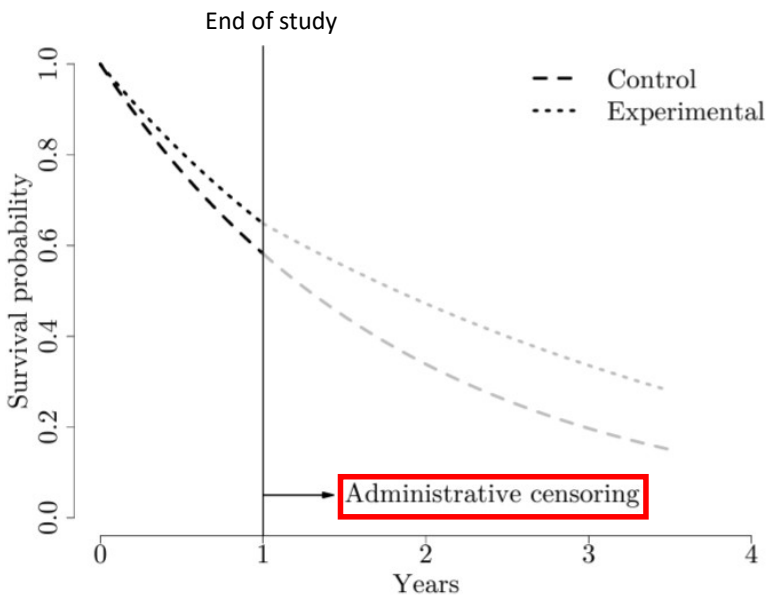
Handling right-censoring

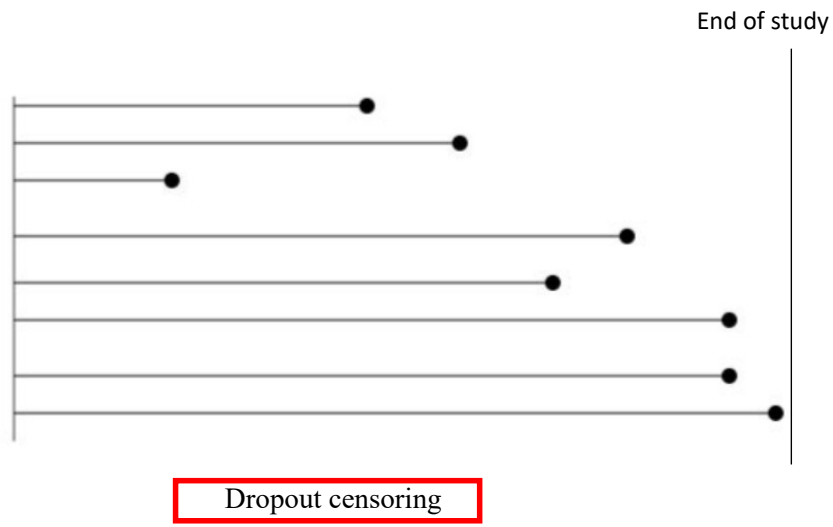
- Drop-out: Peron's scoring rule
- Administrative censoring: restriction time

Right-censoring

We do not observe X and Y

- instead we observe a right-censored version $\tilde{X}$ and $\tilde{Y}$ and corresponding censoring indicators ϵ_X and ϵ_Y.
- administrative censoring at a deterministic τ: $\tilde{X} = X \wedge \tau$
- drop-out at a random C : $\tilde{X} = X \wedge C$

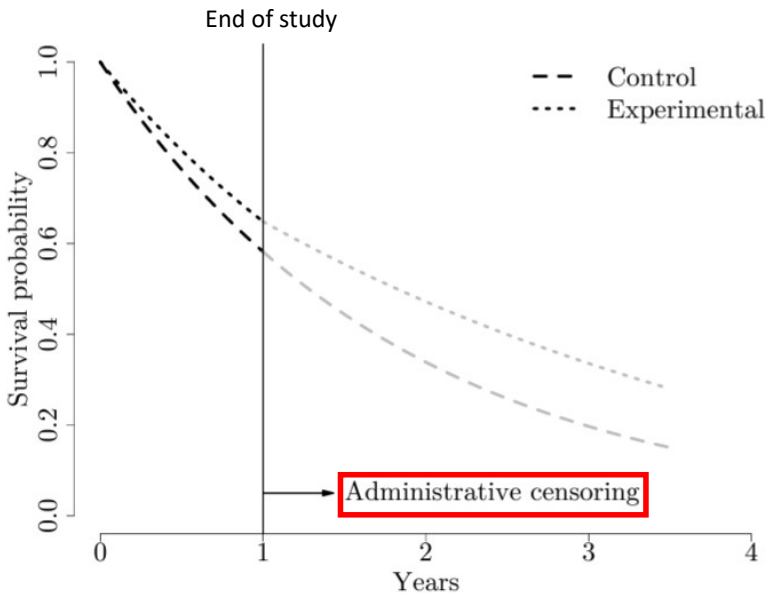


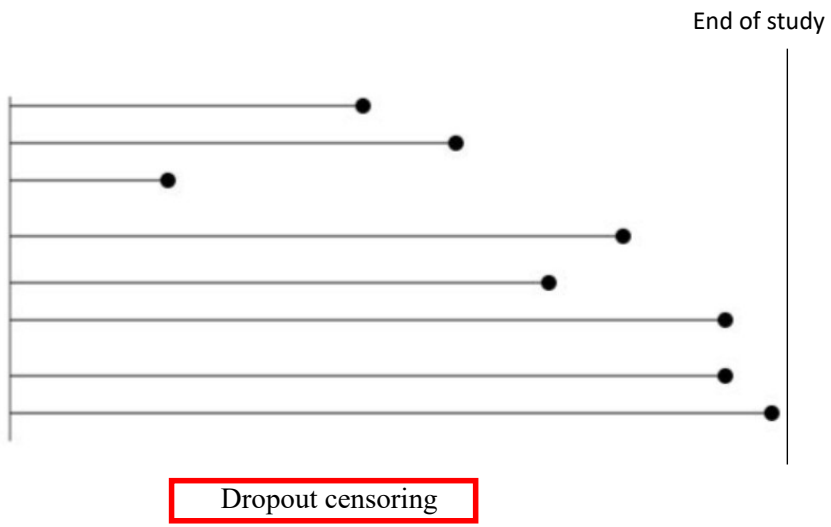


Right-censoring

We do not observe X and Y

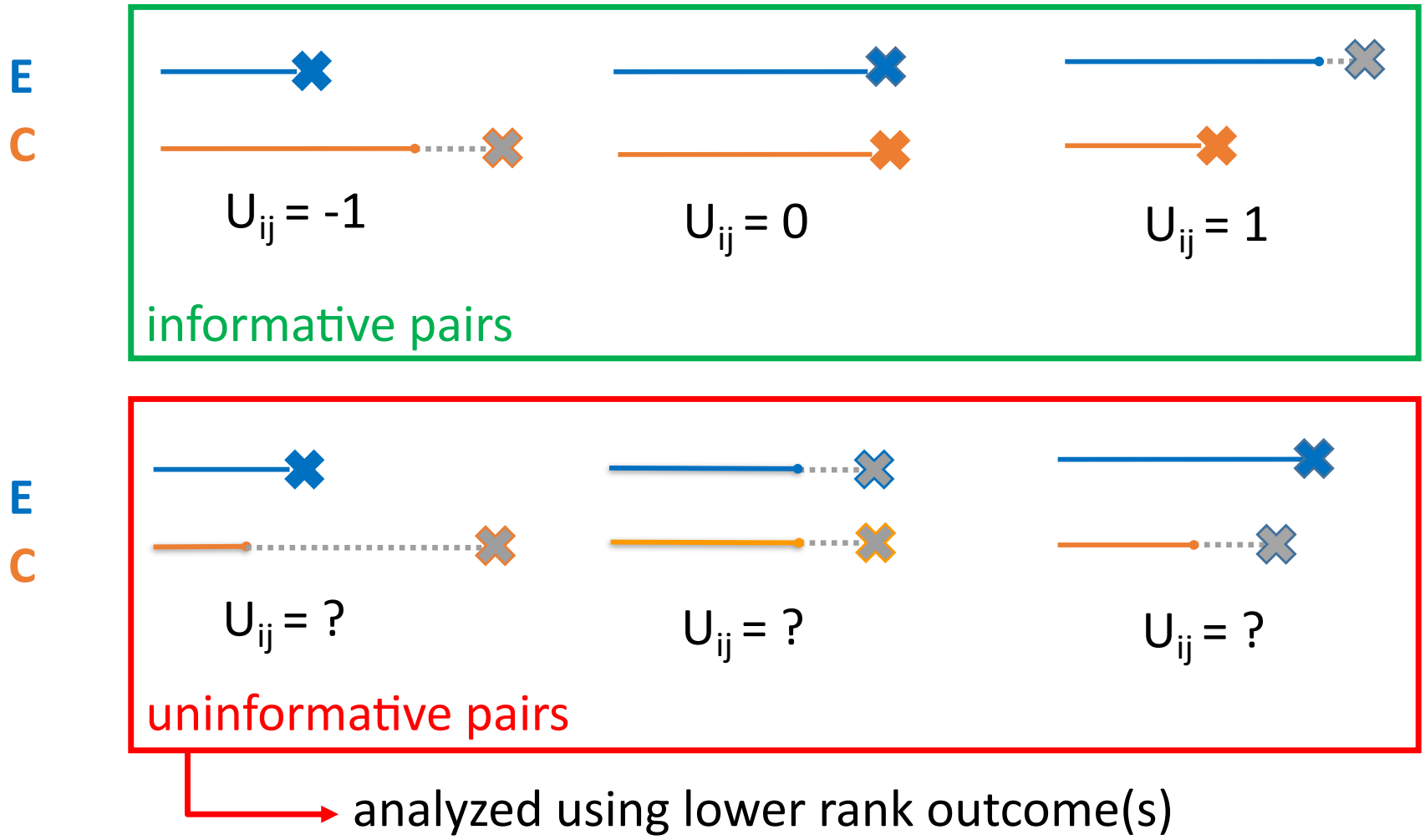
- instead we observe a right-censored version $\tilde{X}$ and $\tilde{Y}$ and corresponding censoring indicators ϵ_X and ϵ_Y.
- administrative censoring at a deterministic τ: $\tilde{X} = X \wedge \tau$
- drop-out at a random C : $\tilde{X} = X \wedge C$



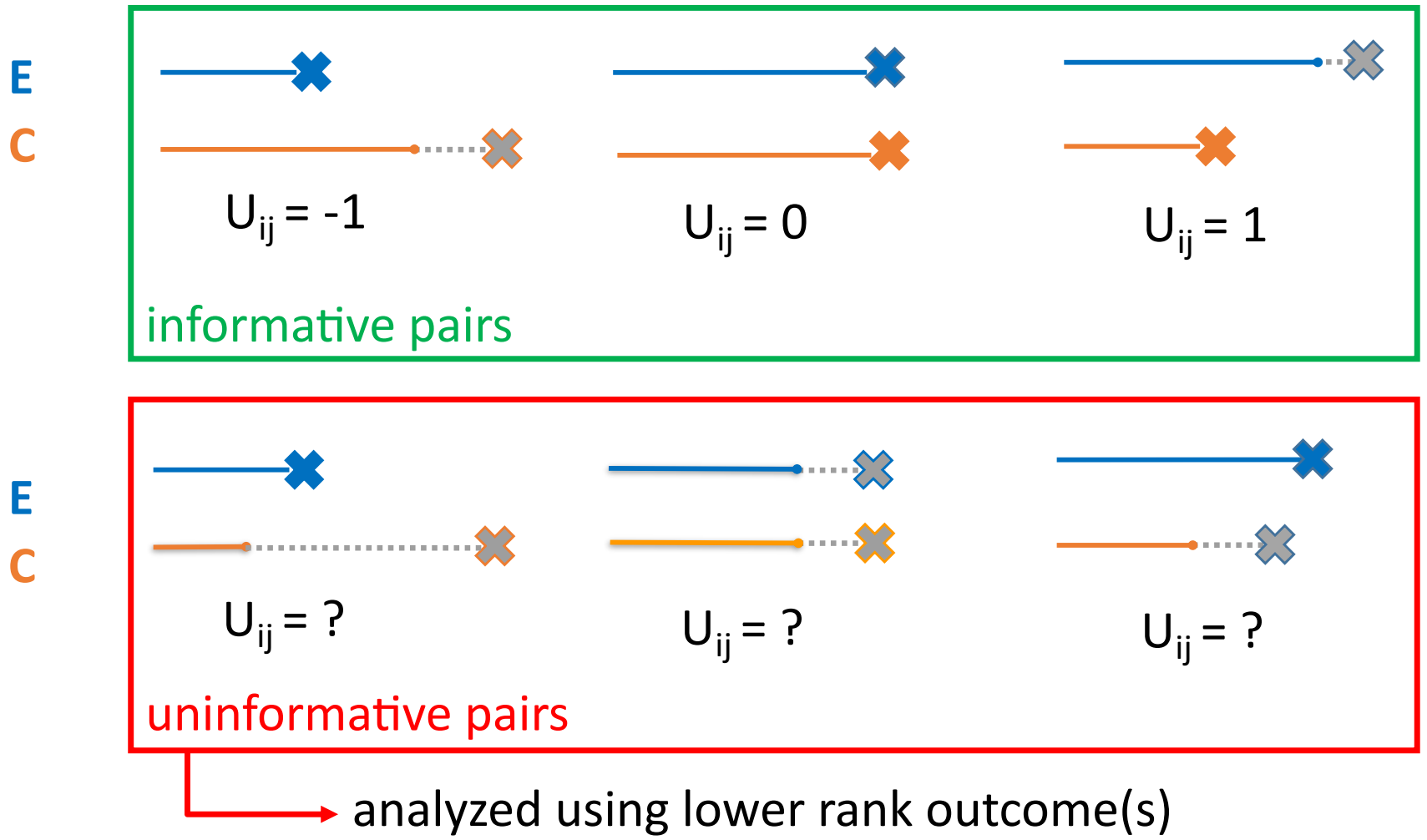


Interest still lies in $\Delta = \mathbb{P}[X \geq Y + \tau] - \mathbb{P}[Y \geq X + \tau]$

Gehan's scoring rule

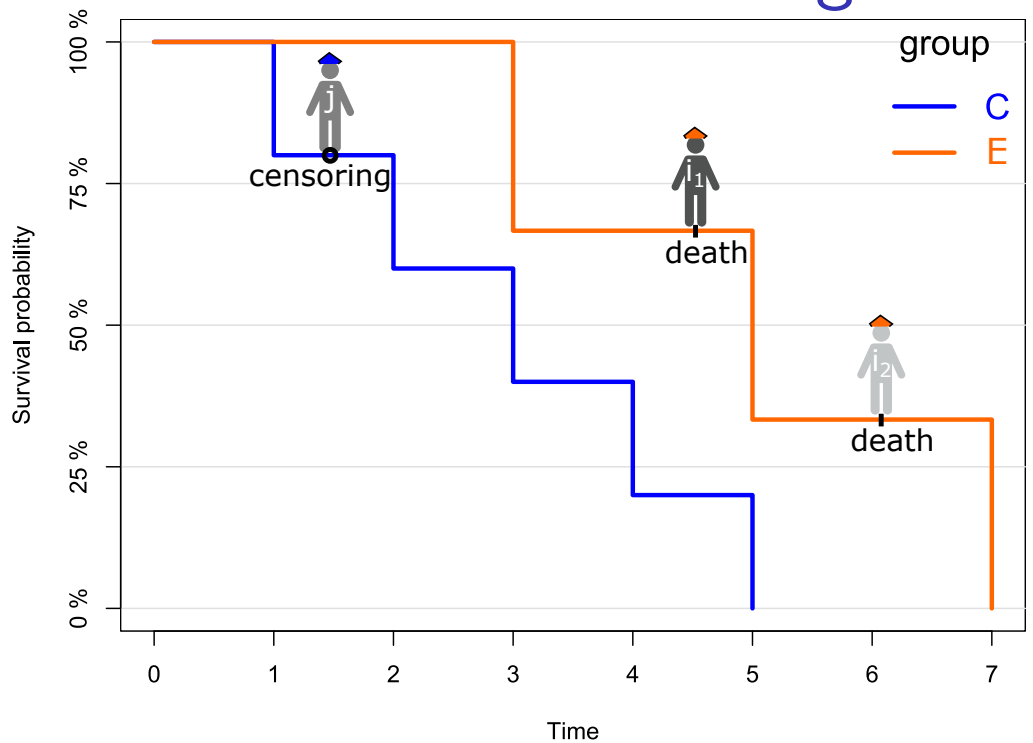


Gehan's scoring rule



⚠ biased towards 0 due uniformative pairs

Peron's scoring rule (example)



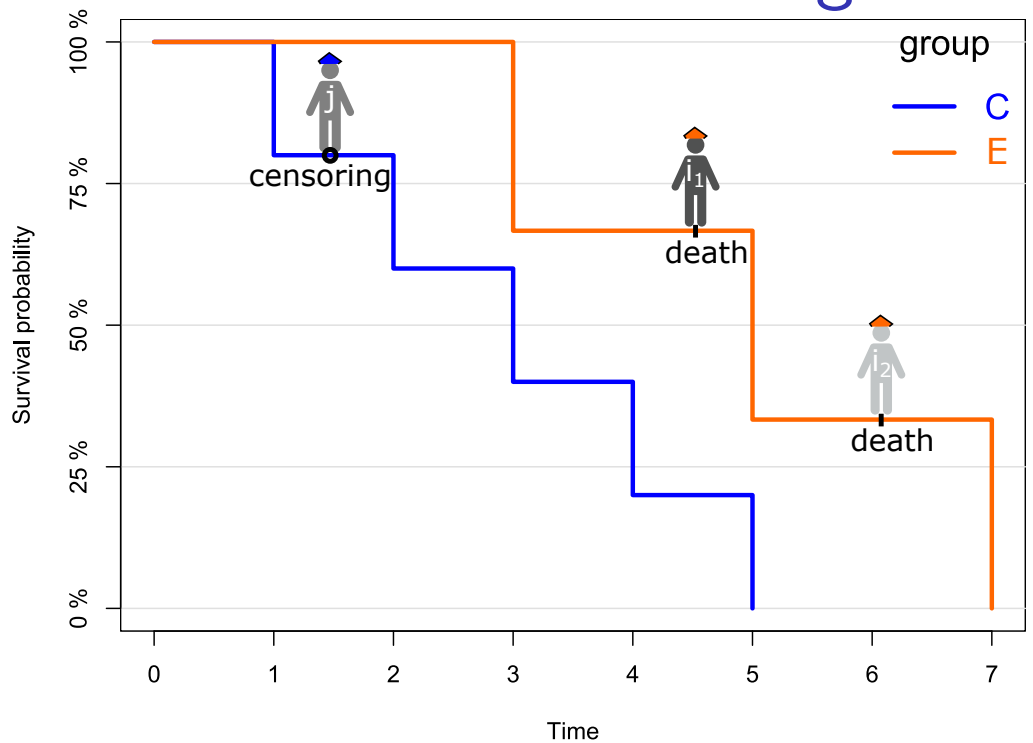
Observed data:

$$(\tilde{x}_{i_1}, \tilde{x}_{i_2}, \tilde{y}_j, \epsilon_{x,i_1}, \epsilon_{x,i_2}, \epsilon_{y,j}) = (4.7, 6.1, 1.5, 1, 1, 0)$$

$\mathbb{P}[X_{i_1} > Y_j | X_{i_1} = \tilde{x}_{i_1}, Y_j > \tilde{y}_j] = 0.75$ Out of the 4 remaining individuals, only 1 survived up to the observed event time

$\mathbb{P}[X_{i_2} > Y_j | X_{i_2} = \tilde{x}_{i_2}, Y_j > \tilde{y}_j] = 1$ no survivor at the censoring time in the other group.

Peron's scoring rule (example)



Observed data:

$$(\tilde{x}_{i_1}, \tilde{x}_{i_2}, \tilde{y}_j, \epsilon_{x,i_1}, \epsilon_{x,i_2}, \epsilon_{y,j}) = (4.7, 6.1, 1.5, 1, 1, 0)$$

Introducing S_X and S_Y the group-specific survival curves:

$$\mathbb{E}\left[\mathbb{1}_{X \geq Y+\tau} | X = \tilde{x}_i, Y > \tilde{y}_j\right] = 1 - \frac{S_Y(x_i - \tau)}{S_Y(\tilde{y}_j)}$$

$$= 1 - \frac{0.2}{0.8} = 0.75 \text{ (for } i_1 \text{ vs. } j \text{ with } \tau = 0.0001)$$

Peron scoring rule

For $\mathbb{P}\left[X > Y + \tau | \tilde{X}, \tilde{Y}, \epsilon_X, \epsilon_Y\right]$:

(ϵ_X, ϵ_Y)	$\tilde{X} \leq \tilde{Y} - \tau$	$\lvert\tilde{X} - \tilde{Y}\rvert < \tau$	$\tilde{X} \geq \tilde{Y} + \tau$
$(1,\ 1)$	0	0	1
$(1,\ 0)$	0	0	$1 - \frac{S_Y(\tilde{X}-\tau)}{S_Y(\tilde{Y})}$
$(0,\ 1)$	$\frac{S_X(\tilde{Y}+\tau)}{S_X(\tilde{X})}$	$\frac{S_X(\tilde{Y}+\tau)}{S_X(\tilde{X})}$	1
$(0,\ 0)$	$-\frac{I(\tilde{Y})}{S_X(\tilde{X})S_Y(\tilde{Y})}$	$-\frac{I(\tilde{Y})}{S_X(\tilde{X})S_Y(\tilde{Y})}$	$1 - \frac{S_Y(\tilde{X}-\tau)}{S_Y(\tilde{Y})} - \frac{I(\tilde{X}-\tau)}{S_X(\tilde{X})S_Y(\tilde{Y})}$

with $I(s) = \int_{t>s}^{\infty} S_X(t+\tau) dS_Y(t)$.

A similar formula holds for $\mathbb{P}\left[Y > X_i + \tau | \tilde{X}_i, \tilde{Y}, \epsilon_X, \epsilon_Y\right]$

package BuyseTest

```
eG.BT <- BuyseTest(treatment ~ tte(OS, statusOS, threshold = 5),
  data = prodige, keep.pairScore = TRUE, scoring.rule = "Gehan")
getPairScore(eG.BT)[c(1,2,169238),]
```

	index.C	index.T	favorable	unfavorable	neutral	uninf	weight
1:	1	403	0	0	0	1	1
2:	2	403	0	0	0	1	1
3:	398	823	1	0	0	0	1

```
eP.BT <- BuyseTest(treatment ~ tte(OS, statusOS, threshold = 5),
  data = prodige, keep.pairScore = TRUE) ## Peron is default
print(getPairScore(eP.BT)[c(1,2,169238),], digits = 2)
```

	index.C	index.T	favorable	unfavorable	neutral	uninf	weight
1:	1	403	0.62	0.00	0.38	0.0000	1
2:	2	403	0.34	0.22	0.43	0.0024	1
3:	398	823	1.00	0.00	0.00	0.0000	1

GPC methods for right-censored outcomes

Naive approaches:

- Gehan: uninformative pairs = 0

Imputation approaches: survival model

- Peron: Kaplan-Meier (KM) stratified on treatment arm

GPC methods for right-censored outcomes

Naive approaches:

- Gehan: uninformative pairs = 0
- Harrell: ignore uninformative pairs (biased except under PH)

Imputation approaches: survival model

- Latta: Kaplan-Meier (KM) common to both arms
- Peron: Kaplan-Meier (KM) 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 (inefficient)
- Dong: 'Gehan'-like alternative

Comments

BuyseTest implements "Gehan", "Peron", "Efron" via the argument `scoring.rule`.

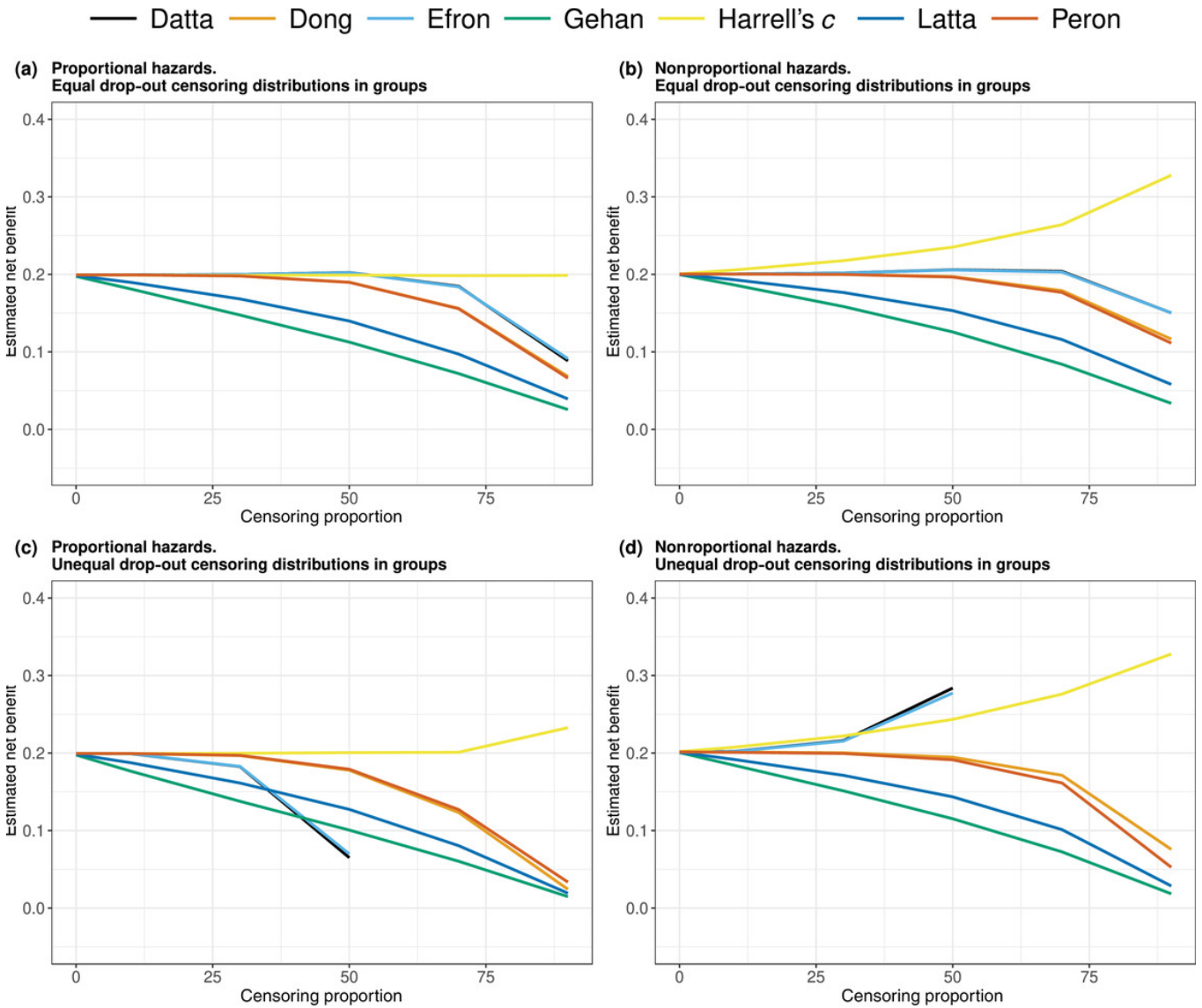
The user can also provide his own survival model (argument `model.tte`)

- provided it is a `prodlim` or `survreg` object
- depending on time, treatment, and strata covariates
- easy to do "Latta", could be tuned to do "De Backer".

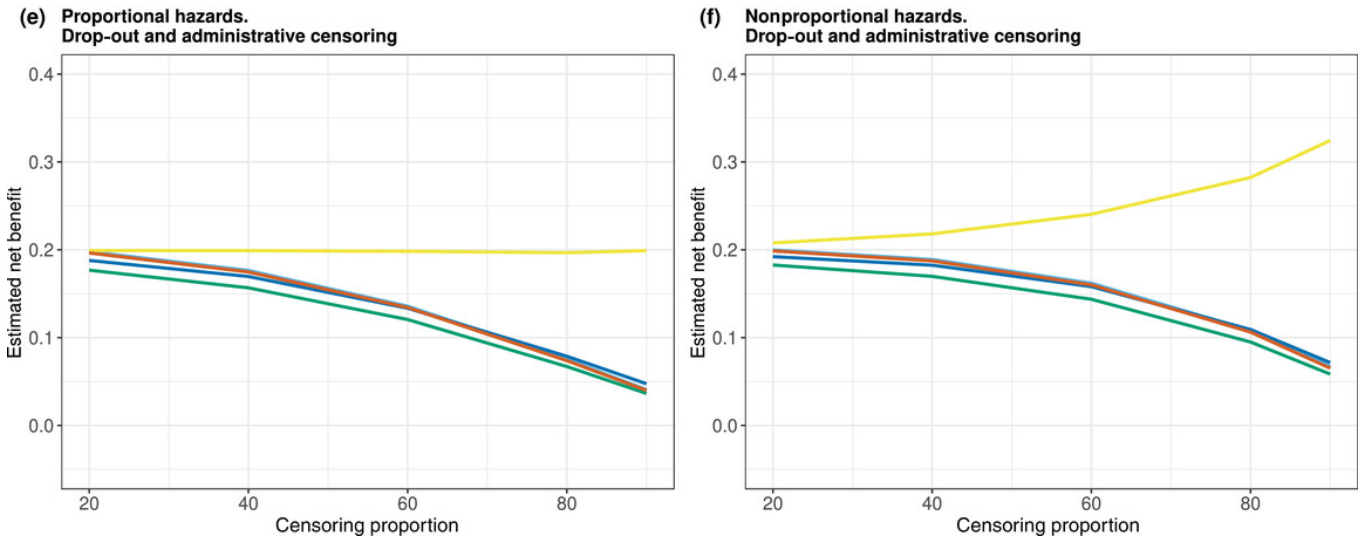
Weighting approach are not implemented:

- relevant with covariates
- otherwise specifying a KM imputation model ($\widehat{S}$) is equivalent to a KM model for censoring ($\widehat{G}$) since $\widehat{S}(t)\widehat{G}(t) = \frac{Y(t)}{n}$ ($Y(t)$ number of subject remaining at risk at time t)

Performance in presence of drop-out



What about administrative censoring?



😱 all estimators are substantially downward biased.

An integral form of the NTB

$$\mathbb{P}[\boldsymbol{X} > \boldsymbol{Y} + \tau] = \int_0^{+\infty} \mathbb{P}[x > \boldsymbol{Y} + \tau]\, dF_X(x) = \int_0^{+\infty} F_Y(x+\tau) dF_X(x)$$

The NTB can be re-written as:

$$\Delta = \int_0^{+\infty} F_Y(x+\tau) dF_X(x) - \int_0^{+\infty} F_X(y+\tau) dF_Y(y)$$

An integral form of the NTB

$$\mathbb{P}[\boldsymbol{X} > \boldsymbol{Y} + \tau] = \int_0^{+\infty} \mathbb{P}[x > \boldsymbol{Y} + \tau]\, dF_X(x) = \int_0^{+\infty} F_Y(x+\tau) dF_X(x)$$

The NTB can be re-written as:

$$\Delta = \int_0^{+\infty} F_Y(x+\tau) dF_X(x) - \int_0^{+\infty} F_X(y+\tau) dF_Y(y)$$

In presence of administrative censoring, say at γ, the tail of the distribution is unknown and so are F_X and F_Y beyond γ.

- we cannot get a non-parametric estimator for Δ

An integral form of the NTB

$$\mathbb{P}[\boldsymbol{X} > \boldsymbol{Y} + \tau] = \int_0^{+\infty} \mathbb{P}[x > \boldsymbol{Y} + \tau]\, dF_X(x) = \int_0^{+\infty} F_Y(x+\tau) dF_X(x)$$

The NTB can be re-written as:

$$\Delta = \int_0^{+\infty} F_Y(x+\tau) dF_X(x) - \int_0^{+\infty} F_X(y+\tau) dF_Y(y)$$

In presence of administrative censoring, say at γ, the tail of the distribution is unknown and so are F_X and F_Y beyond γ.

- we cannot get a non-parametric estimator for Δ
- Peron's scoring rule can be understood at stopping the integration at the last observation in each group $(X_{\max}, Y_{\max})$

$$\Delta = \int_0^{X_{\max} \wedge (Y_{\max} - \tau)} F_Y(x+\tau) dF_X(x) - \int_0^{Y_{\max} \wedge (X_{\max} - \tau)} F_X(y+\tau) dF_Y(y)$$

Proportion of uninformative pairs

The R package BusyeTest keeps track of the amount 'missing' from the two integrals (`uninf`):

```
print(model.tables(e.BT)[,1:7], digits = 3)
```

```
  endpoint threshold total favorable unfavorable neutral  uninf
1       OS     5e+00 100.0      30.8        18.0    51.1 0.0793
3 toxicity     1e-12  51.2      17.4        19.2    14.6 0.0000
```

Administrative censoring: solution(s)

Make parametric assumptions:

- De Backer et al. (2023): tail of the survival curve follow an extreme value distribution.
- Péron et al. (2021): 'redistribute the tail' under proportional hazard

Change estimand: (recommended)

- Péron et al. (2021): use the restricted NTB

$$\Delta = \int_0^{\gamma-\tau} F_Y(x+\tau)dF_X(x) - \int_0^{\gamma-\tau} F_X(y+\tau)dF_Y(y)$$
$$= \mathbb{P}\left[\boldsymbol{X} \wedge \gamma > \boldsymbol{Y} \wedge \gamma + \tau\right] - \mathbb{P}\left[\boldsymbol{Y} \wedge \gamma > \boldsymbol{X} \wedge \gamma + \tau\right]$$

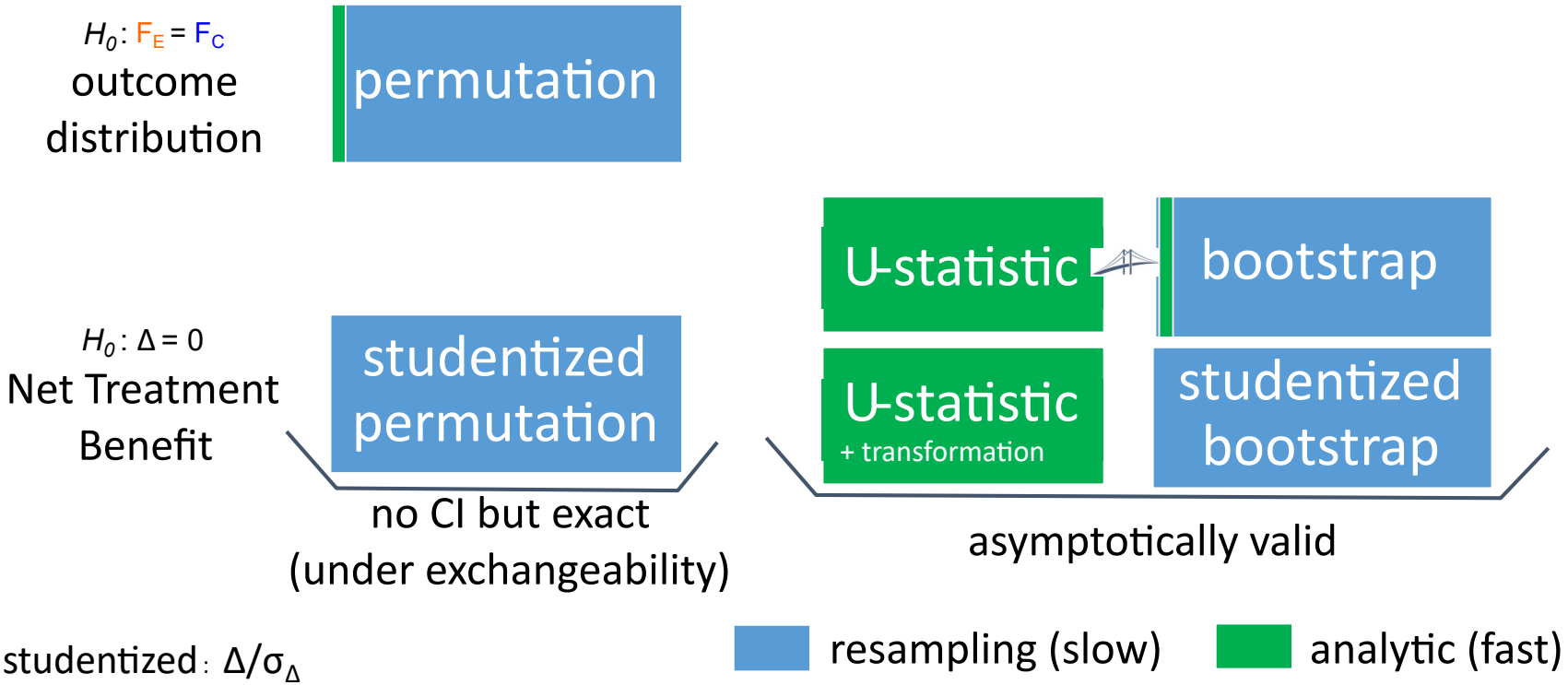
```
BuyseTest(treatment ~ tte(OS, statusOS, threshold = 5, restriction =
    30), data = prodige)
```

favorable	unfavorable	neutral	uninf
30.61	17.98	51.41	0.00

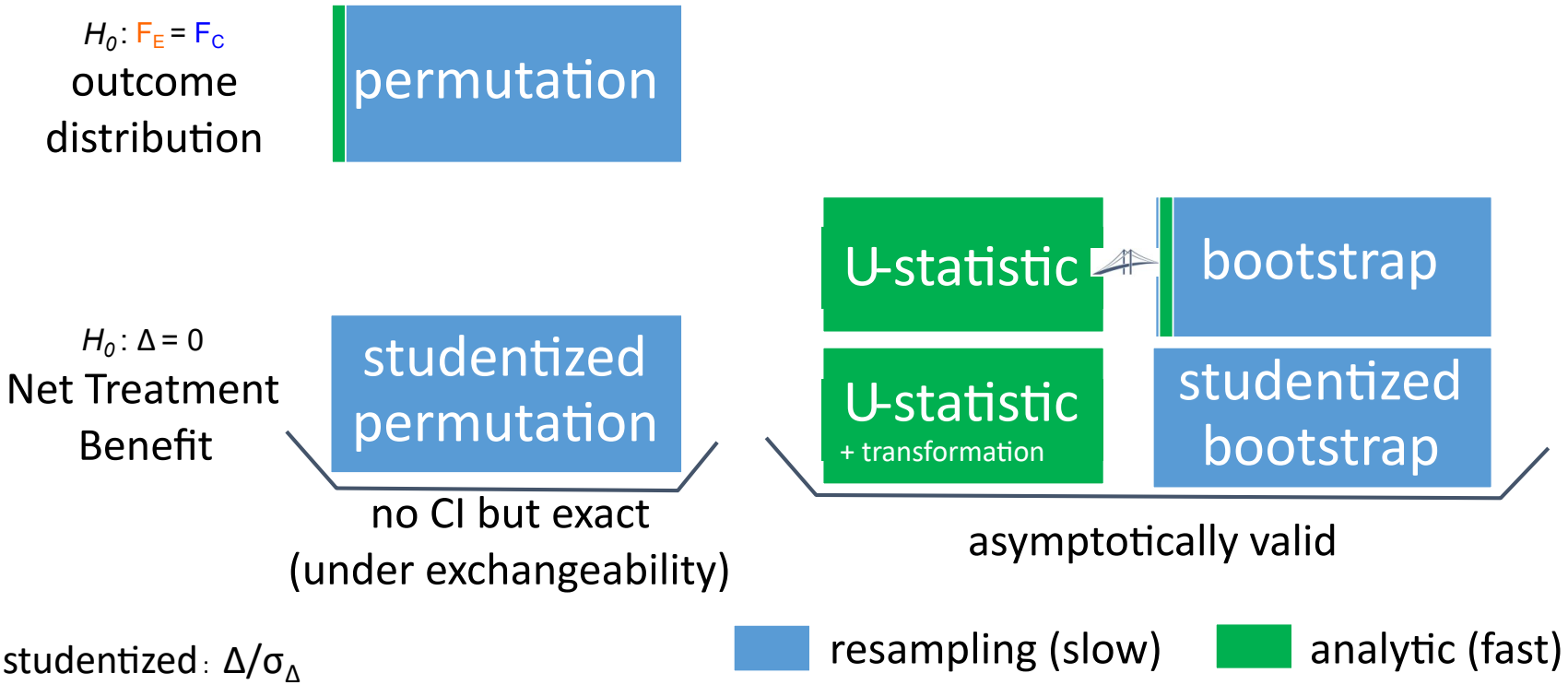
Statistical inference

- Overview
  - U-statistic approach

Overview of method implemented in BuyseTest



Overview of method implemented in BuyseTest



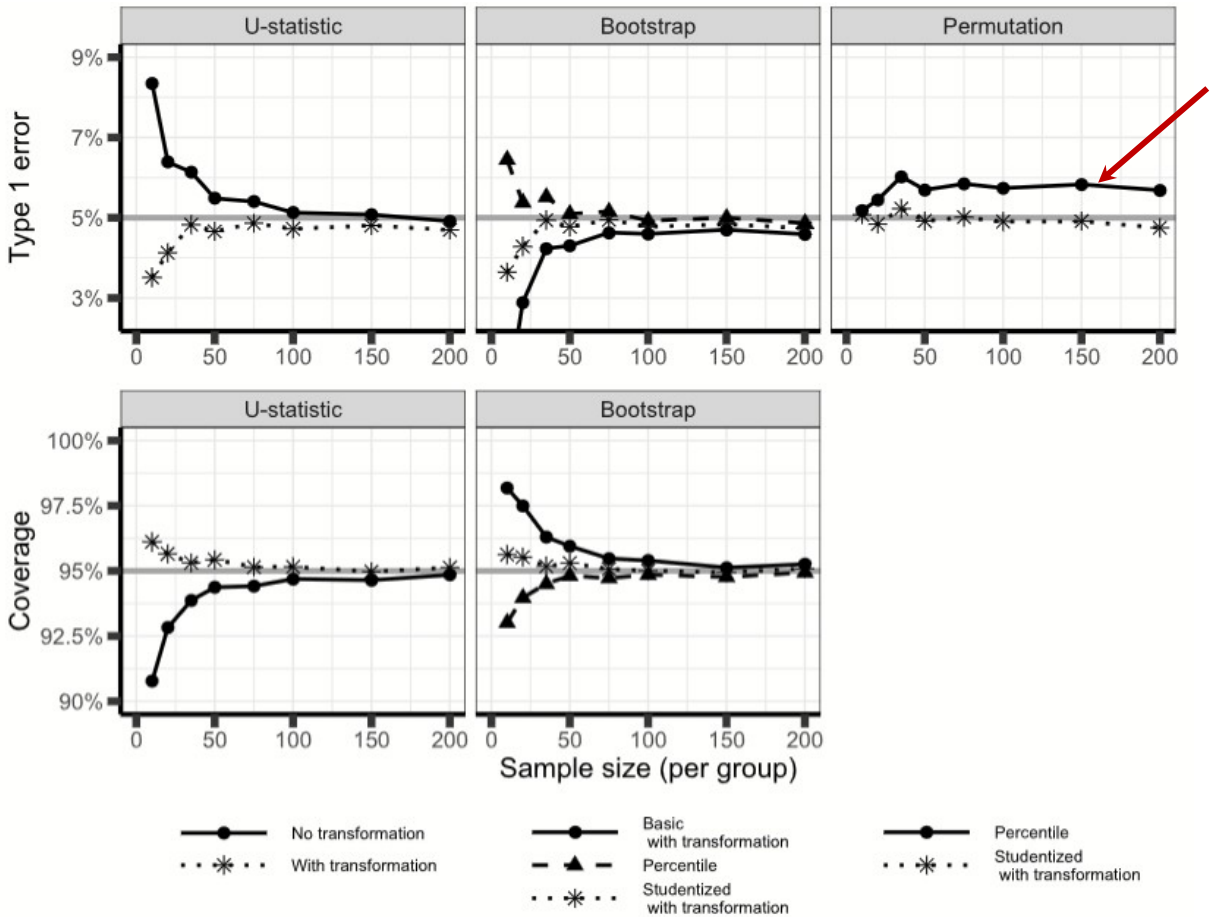
Analytic bootstrap and permutation (Anderson and Verbeeck, 2023)

- analytic variance estimator, normally distributed estimate.
- bootstrap same as 2nd order U-statistic

Results from a simulation study

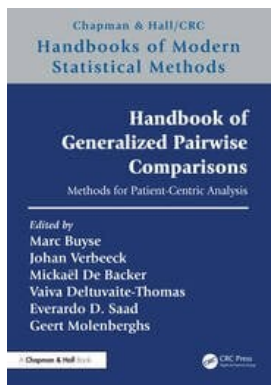
Setting: normally distributed outcome with variance 2 or 1.

- mean difference: 0 (type 1 error) or 1 (coverage)



heteroscedastic outcome

From

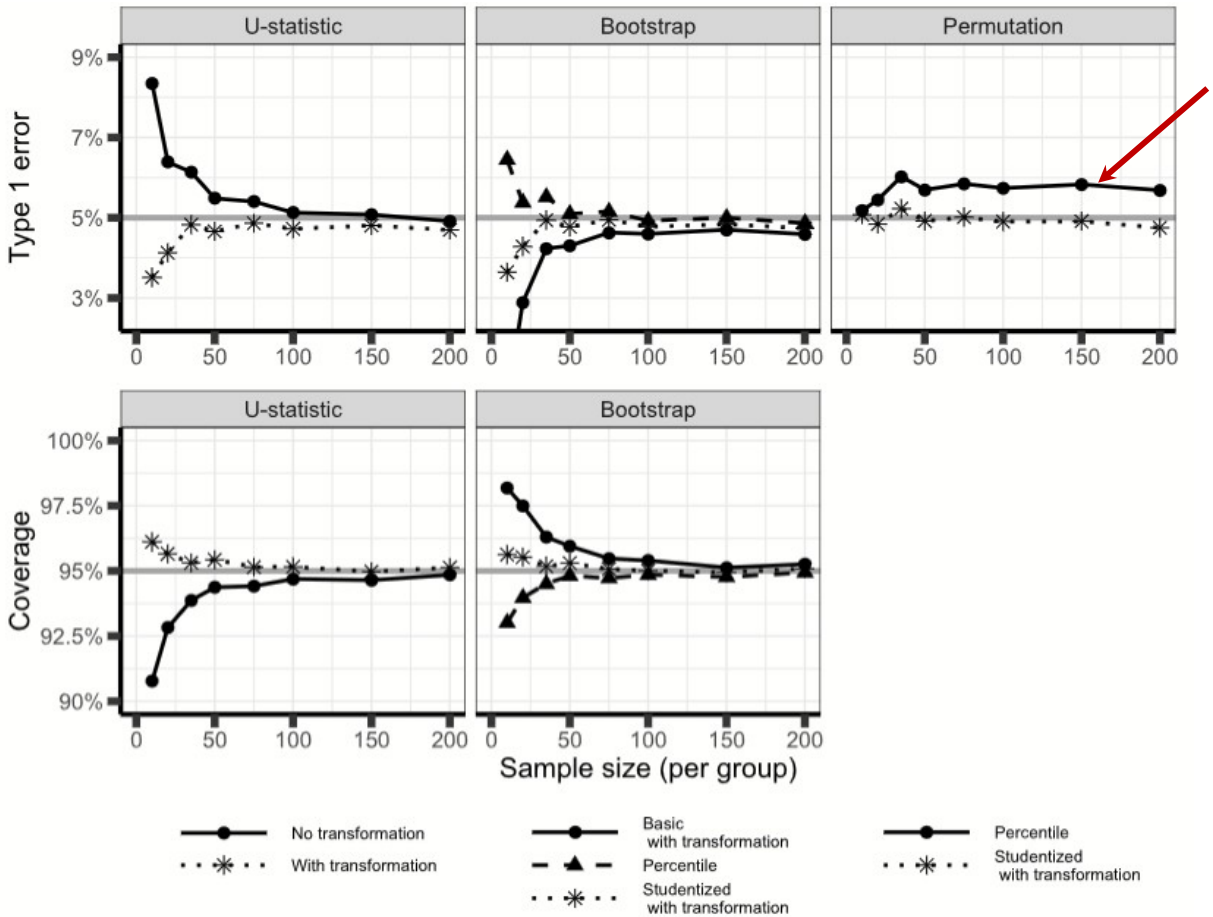


chapter 3

Results from a simulation study

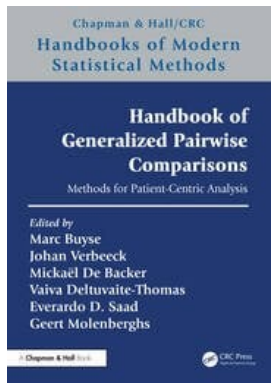
Setting: normally distributed outcome with variance 2 or 1.

- mean difference: 0 (type 1 error) or 1 (coverage)



heteroscedastic outcome

From



chapter 3

😱 the 'usual' Wilcoxon test rejects more than 5%

U-statistic theory - Intuition

$$\widehat{\Delta} = \frac{1}{n_X n_Y} \sum_{i=1}^{n_X} \sum_{j=1}^{n_Y} \mathbb{1}_{X_i \geq Y_j + \tau} - \mathbb{1}_{Y_j \geq Y_i + \tau} \text{ is an average !}$$

U-statistic theory - Intuition

$$\widehat{\Delta} = \frac{1}{n_X n_Y} \sum_{i=1}^{n_X} \sum_{j=1}^{n_Y} \mathbb{1}_{X_i \geq Y_j + \tau} - \mathbb{1}_{Y_j \geq Y_i + \tau} \text{ is an average !}$$

⚠ $\left(\mathbb{1}_{X_i \geq Y_j + \tau} - \mathbb{1}_{Y_j \geq X_i + \tau} \right)_{i \in \{1,\dots,n\}, j \in \{1,\dots,m\}}$ are not independent

$\mathbb{1}_{X_i \geq Y_j + \tau} - \mathbb{1}_{Y_j \geq X_i + \tau}$ and $\mathbb{1}_{X_i \geq Y_{j'} + \tau} - \mathbb{1}_{Y_{j'} \geq X_i + \tau}$ both depends on X_i

U-statistic theory - Intuition

$$\widehat{\Delta} = \frac{1}{n_X n_Y} \sum_{i=1}^{n_X} \sum_{j=1}^{n_Y} \mathbb{1}_{X_i \geq Y_j + \tau} - \mathbb{1}_{Y_j \geq Y_i + \tau} \text{ is an average !}$$

⚠ $\left(\mathbb{1}_{X_i \geq Y_j + \tau} - \mathbb{1}_{Y_j \geq X_i + \tau} \right)_{i \in \{1,\dots,n\}, j \in \{1,\dots,m\}}$ are not independent

$\mathbb{1}_{X_i \geq Y_j + \tau} - \mathbb{1}_{Y_j \geq X_i + \tau}$ and $\mathbb{1}_{X_i \geq Y_{j'} + \tau} - \mathbb{1}_{Y_{j'} \geq X_i + \tau}$ both depends on X_i

This motivate the following H-decomposition (Lee, 1990):

$$\widehat{\Delta} - \Delta = \frac{1}{n_X} \sum_{i=1}^{n_X} H_i^{(1,0)} + \frac{1}{n_Y} \sum_{j=1}^{n_Y} H_j^{(0,1)} + \frac{1}{n_X n_Y} \sum_{i=1}^{n_X} \sum_{j=1}^{n_Y} H_{ij}^{(1,1)}$$

- sum of uncorrelated U-statistics of increasing order
- with variance of decreasing order in n_X, n_Y.

Back to the Central Limit Theorm (CLT)

$$\widehat{\Delta} - \Delta = \frac{1}{n_X} \sum_{i=1}^{n_X} H_i^{(1,0)} + \frac{1}{n_Y} \sum_{j=1}^{n_Y} H_j^{(0,1)} + \underbrace{\frac{1}{n_X n_Y} \sum_{i=1}^{n_X} \sum_{j=1}^{n_Y} H_{ij}^{(1,1)}}_{\text{asymptotically neglectable}}$$

$$H_i^{(1,0)} = \mathbb{E}[\mathbb{1}_{X_i \geq Y_j + \tau} - \mathbb{1}_{Y_j \geq X_i + \tau} | X_i] - \Delta$$

$$H_j^{(0,1)} = \mathbb{E}[\mathbb{1}_{X_i \geq Y_j + \tau} - \mathbb{1}_{Y_j \geq X_i + \tau} | Y_j] - \Delta$$

Back to the Central Limit Theorm (CLT)

$$\widehat{\Delta} - \Delta = \frac{1}{n_X}\sum_{i=1}^{n_X} H_i^{(1,0)} + \frac{1}{n_Y}\sum_{j=1}^{n_Y} H_j^{(0,1)} + \underbrace{\frac{1}{n_X n_Y}\sum_{i=1}^{n_X}\sum_{j=1}^{n_Y} H_{ij}^{(1,1)}}_{\text{asymptotically neglectable}}$$

$$H_i^{(1,0)} = \mathbb{E}[\mathbb{1}_{X_i \geq Y_j + \tau} - \mathbb{1}_{Y_j \geq X_i + \tau} | X_i] - \Delta$$

$$H_j^{(0,1)} = \mathbb{E}[\mathbb{1}_{X_i \geq Y_j + \tau} - \mathbb{1}_{Y_j \geq X_i + \tau} | Y_j] - \Delta$$

- CLT: $\frac{1}{n_X}\sum_{i=1}^{n_X} H_i^{(1,0)}$ and $\frac{1}{n_Y}\sum_{j=1}^{n_Y} H_j^{(0,1)}$ are asymptotically normally distributed.
- using independence between groups, $\widehat{\Delta}$ is also asymptotically normally distributed.

$$\mathbb{V}ar\left[\widehat{\Delta}\right] \approx \frac{1}{n_X^2}\sum_{i=1}^{n_X}\left(H_i^{(1,0)}\right)^2 + \frac{1}{n_Y^2}\sum_{j=1}^{n_Y}\left(H_j^{(0,1)}\right)^2$$

Example

Control observations

−1.2 −0.5 −0.8 0.3 1.1 1.2 0.7 −0.5 0.6 −1.2

Experimental observations

−0.6
−2.2
−0.7
−2.1
−1.3
−0.4
−0.7
−0.9
−0.1
−0.3

Example

Control observations

Experimental observations	−1.2	−0.5	−0.8	0.3	1.1	1.2	0.7	−0.5	0.6	−1.2
−0.6	1	−1	1	−1	−1	−1	−1	−1	−1	1
−2.2	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1
−0.7	1	−1	1	−1	−1	−1	−1	−1	−1	1
−2.1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1
−1.3	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1
−0.4	1	1	1	−1	−1	−1	−1	1	−1	1
−0.7	1	−1	1	−1	−1	−1	−1	−1	−1	1
−0.9	1	−1	−1	−1	−1	−1	−1	−1	−1	1
−0.1	1	1	1	−1	−1	−1	−1	1	−1	1
−0.3	1	1	1	−1	−1	−1	−1	1	−1	1

Example

Control observations

Experimental observations	−1.2	−0.5	−0.8	0.3	1.1	1.2	0.7	−0.5	0.6	−1.2	$U_{i,\bullet}^{+}$	$U_{i,\bullet}^{-}$	$U_{i,\bullet}^{+}-U_{i,\bullet}^{-}$	$\left(\frac{U_{i,\bullet}^{+}-U_{i,\bullet}^{-}}{n_Y}-\widehat{\Delta}\right)^2$
−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
											+			
											26	74	-48	1.536

$$\widehat{\Delta} = \frac{U_{\bullet,\bullet}^{+} - U_{\bullet,\bullet}^{-}}{n_X n_Y} = \frac{26 - 74}{100} = -0.48$$

$$\mathbb{V}ar\left[\widehat{\Delta}\right] \approx \frac{1}{n_X^2}\sum_{i=1}^{n_X}\left(H_i^{(1,0)}\right)^2 + \frac{1}{n_Y^2}\sum_{j=1}^{n_Y}\left(H_j^{(0,1)}\right)^2 = \frac{1.536}{100} + \frac{?.???}{100}$$

Example

Control observations

Experimental observations	−1.2	−0.5	−0.8	0.3	1.1	1.2	0.7	−0.5	0.6	−1.2	$U_{i,\bullet}^{+}$	$U_{i,\bullet}^{-}$	$U_{i,\bullet}^{+}-U_{i,\bullet}^{-}$	$\left(\frac{U_{i,\bullet}^{+}-U_{i,\bullet}^{-}}{n_Y}-\widehat{\Delta}\right)^2$
−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
											+			
$U_{\bullet,j}^{+}-U_{\bullet,j}^{-}$	4	-4	2	-10	-10	-10	-10	-4	-10	4	26	74	-48	1.536
$\left(\frac{U_{\bullet,j}^{+}-U_{\bullet,j}^{-}}{n_X}-\widehat{\Delta}\right)^2$	0.7744	0.0064	0.4624	0.2704	0.2704	0.2704	0.2704	0.0064	0.2704	0.7744	+			3.376

$$\widehat{\Delta} = \frac{U_{\bullet,\bullet}^{+} - U_{\bullet,\bullet}^{-}}{n_X n_Y} = \frac{26 - 74}{100} = -0.48$$

$$\mathbb{V}ar\left[\widehat{\Delta}\right] \approx \frac{1}{n_X^2}\sum_{i=1}^{n_X}\left(H_i^{(1,0)}\right)^2 + \frac{1}{n_Y^2}\sum_{j=1}^{n_Y}\left(H_j^{(0,1)}\right)^2 = \frac{1.536}{100} + \frac{3.376}{100} = 0.0491$$

R package BuyseTest

U-statistic + atanh transformation is the default.

```
NTB.iid <- getIid(e.BT) ## H_i/n_X or H_j/n_Y
NTB.iid[1:3,]
```

```
             OS_t5      toxicity
[1,]  0.0007489663  3.310607e-05
[2,] -0.0006312873 -1.312940e-04
[3,]  0.0007858605  8.179338e-05
```

```
## on the original scale (before atanh transformation)
rbind(meanIid = colMeans(NTB.iid),
      sqrtSSIid = sqrt(colSums(NTB.iid^2)),
      se = confint(e.BT)$se)
```

```
                OS_t5      toxicity
meanIid   -2.815379e-17 -5.196513e-17
sqrtSSIid  3.683740e-02  4.122695e-02
se         3.683740e-02  4.122695e-02
```

Conclusion

GPC in a nutshell

Principled way to combine outcomes

- hierarchy, threshold, restriction time require careful considerations
- patient centric: what treatment benefits most the patient
  Alternative estimand for 'non-inferiority trials' (Backer et al., 2024)
- transparent: contribution of each outcome to the NTB

A 'mature' framework

- right-censoring, competing risks, covariates, multiple testing
- R package BuyseTest attempts to provide a unified interface

despite remaining open questions

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

Comments or questions?

Reference I

Anderson, W. N. and Verbeeck, J. (2023). Exact permutation and bootstrap distribution of generalized pairwise comparisons statistics. *Mathematics*, 11(6):1502.

Backer, M. D., Sengar, M., Mathews, V., Salvaggio, S., Deltuvaite-Thomas, V., Chiêm, J.-C., Saad, E. D., and Buyse, M. (2024). Design of a clinical trial using generalized pairwise comparisons to test a less intensive treatment regimen. *Clinical Trials*, 21(2):180–188.

Buyse, M., Verbeeck, J., Saad, E. D., De Backer, M., Deltuvaite-Thomas, V., and Molenberghs, G. (2025). *Handbook of generalized pairwise comparisons: methods for patient-centric analysis*. CRC Press.

Reference II

De Backer, M., Legrand, C., Péron, J., Lambert, A., and Buyse, M. (2023). On the use of extreme value tail modeling for generalized pairwise comparisons with censored outcomes. *Pharmaceutical statistics*, 22(2):284–299.

Fay, M. P., Brittain, E. H., Shih, J. H., Follmann, D. A., and Gabriel, E. E. (2018). Causal estimands and confidence intervals associated with wilcoxon-mann-whitney tests in randomized experiments. *Statistics in medicine*, 37(20):2923–2937.

Lee, A. J. (1990). *U-statistics: Theory and Practice*. Routledge.

Moore, M. J., Goldstein, D., Hamm, J., Figer, A., Hecht, J. R., Gallinger, S., Au, H. J., Murawa, P., Walde, D., Wolff, R. A., et al. (2007). Erlotinib plus gemcitabine compared with gemcitabine alone in patients with advanced pancreatic cancer: a phase iii trial of the national cancer institute of canada clinical trials group. *Journal of clinical oncology*, 25(15):1960–1966.

Reference III

Péron, J., Idlhaj, M., Maucort-Boulch, D., Giai, J., Roy, P., Collette, L., Buyse, M., and Ozenne, B. (2021). Correcting the bias of the net benefit estimator due to right-censored observations. *Biometrical Journal*, 63(4):893–906.

Von Hoff, D. D., Ervin, T., Arena, F. P., Chiorean, E. G., Infante, J., Moore, M., Seay, T., Tjulandin, S. A., Ma, W. W., Saleh, M. N., et al. (2013). Increased survival in pancreatic cancer with nab-paclitaxel plus gemcitabine. *New England journal of medicine*, 369(18):1691–1703.

Causal estimand behind PI (1/2)

$$\begin{aligned}\Theta &= \mathbb{P}[\boldsymbol{X}_i > \boldsymbol{Y}_j] = 0.5(2 * \mathbb{P}[\boldsymbol{X}_i > \boldsymbol{Y}_j] - 1) + 0.5 \\ &= 0.5(\mathbb{P}[\boldsymbol{X}_i > \boldsymbol{Y}_j] - \mathbb{P}[\boldsymbol{Y}_j > \boldsymbol{X}_i]) + 0.5\end{aligned}$$

using $1 = \mathbb{P}[\boldsymbol{X}_i > \boldsymbol{Y}_j] + \mathbb{P}[\boldsymbol{Y}_j > \boldsymbol{X}_i]$ with a continuous outcome.

$$\begin{aligned}\Theta &= \mathbb{P}[\boldsymbol{X}_i > \boldsymbol{Y}_j] \\ &= \int_{-\infty}^{+\infty} \mathbb{P}[x > \boldsymbol{Y}_j]\, dF_X(x) = \int_{-\infty}^{+\infty} F_Y(x) dF_X(x) \\ &= \int_{-\infty}^{+\infty} (F_Y(x) + F_X(x)) dF_X(x) - \int_{-\infty}^{+\infty} F_X(x) dF_X(x)\end{aligned}$$

The last term can be re-written $\mathbb{E}[F_X(X)]$ which equals 0.5. Indeed, $F_X(X)$ has standard uniform distribution by the probability integral transform so its expectation is 0.5.

Causal estimand behind PI (2/2)

Introducing:

- $\overline{F}(.) = 0.5F_X(.) + 0.5F_Y(.)$ the 'average' outcome distribution pooling the two potential outcomes distribution with equal weights.

$$\begin{aligned}\Theta &= \int_{-\infty}^{+\infty} (F_Y(x) + F_X(x))dF_X(x) - 0.5 \\ &= 2\int_{-\infty}^{+\infty} \overline{F}(x)dF_X(x) - 0.5 = 2\mathbb{E}\left[\overline{F}(X)\right] - 0.5\end{aligned}$$

Injecting in:

$$\begin{aligned}\Theta &= \mathbb{P}\left[\boldsymbol{X}_i > \boldsymbol{Y}_j\right] = 0.5(\mathbb{P}\left[\boldsymbol{X}_i > \boldsymbol{Y}_j\right] - \mathbb{P}\left[\boldsymbol{Y}_j > \boldsymbol{X}_i\right]) + 0.5 \\ &= \mathbb{E}\left[\overline{F}(X_k)\right] - \mathbb{E}\left[\overline{F}(Y_k)\right] + 0.5\end{aligned}$$