

# Composite endpoints including patient relevant endpoints (Quality of Life)



Interuniversity Institute for Biostatistics  
and statistical Bioinformatics

6 May 2022

Johan Verbeeck

[johan.verbeeck@uhasselt.be](mailto:johan.verbeeck@uhasselt.be)

Data Science Institute/I-Biostat

UHasselt - Belgium



# Content

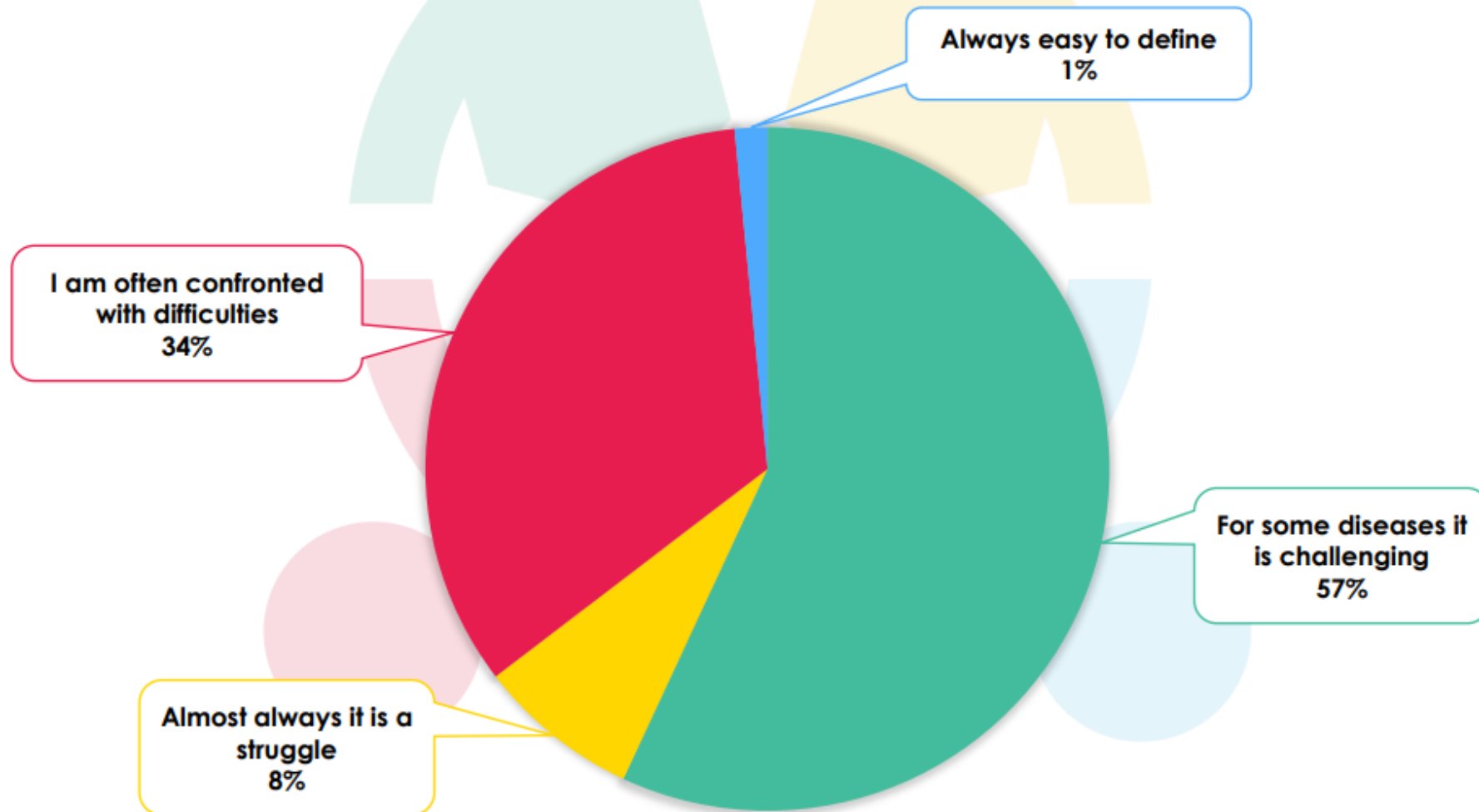
- Composite endpoints in clinical trials
- Generalized Pairwise Comparisons (GPC)
  - Effect measure
  - Characteristics
  - Inference for small samples
- Example: Epidermolysis bullosa
- Conclusions

# Composite endpoints in clinical trials

# Multivariate endpoints

- International Conference Council on Harmonisation recommends to select a **single meaningful endpoint**.

# What is your experience to define a single meaningful endpoint for the study of a disease?



# Multivariate endpoints

- It is **not always easy to choose** or define a meaningful single endpoint
- A single endpoint is **often not sufficient** to reflect the full clinical benefit of a treatment in multifaceted diseases
- **Combination** of several clinical meaningful endpoints

Combination of endpoints of different data type in small sample trials

# Multivariate endpoints methodologies

Combining endpoints on:

- **subject level:**

- Reduce per subject multivariate to univariate endpoint: f.e. clinical indices, composite endpoints (time to first event)

Cox proportional hazard model<sup>1</sup> and its extensions (Anderson-Gill<sup>2</sup>,...)

logrank test and its extension (weighted composite endpoint<sup>3</sup>)

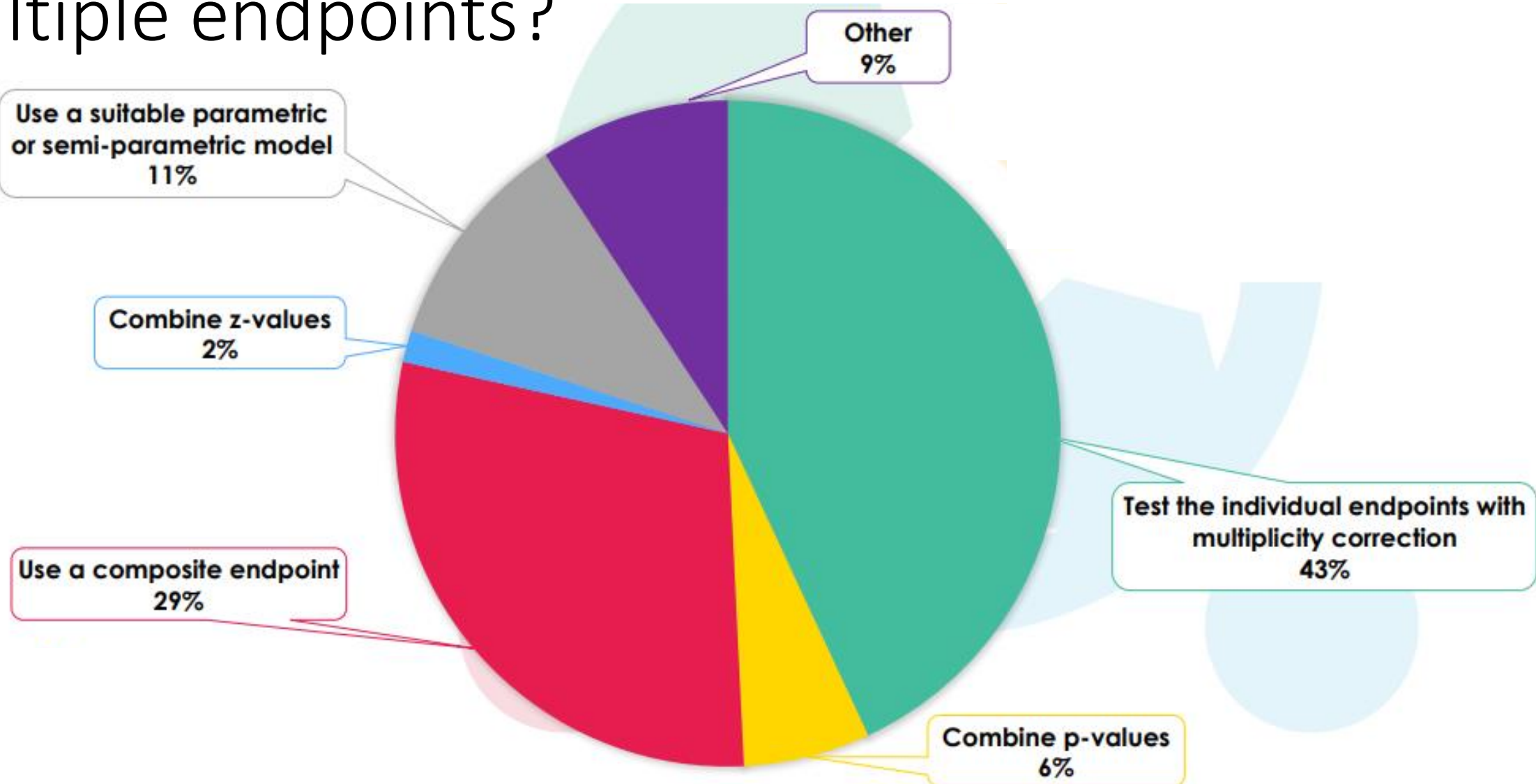
- Joint frailty models<sup>4</sup>,...

are limited:

- In the number and type of endpoints that can be combined
- Poor small sample properties

1. Cox (1972)
2. Andersen and Gill (1982)
3. Armstrong et al. (2011)
4. Rondeau et al. (2007)

# What is your preferred method to handle multiple endpoints?





# Multivariate endpoints methodologies

Combining endpoints on:

- **subject level:**
  - Reduce per subject multivariate to univariate endpoint: f.e. clinical indices, composite endpoints
  - joint models
- **test statistics level:** Combine univariate z-or t-statistics
  - combine t-statistics<sup>1</sup>: accounts for correlation, but only allows for continuous endpoints
  - average z-scores<sup>2</sup>: allows all types of endpoints, but ignores correlation
- **level of p-values:** Combine p-values of endpoint corresponding test f.e. Lancaster<sup>3</sup>, Dai<sup>4</sup> procedures, multiple testing procedures<sup>5</sup> : correlation?

1. O'Brien (1984)

2. Sun et al. (2012)

3. Lancaster (1961)

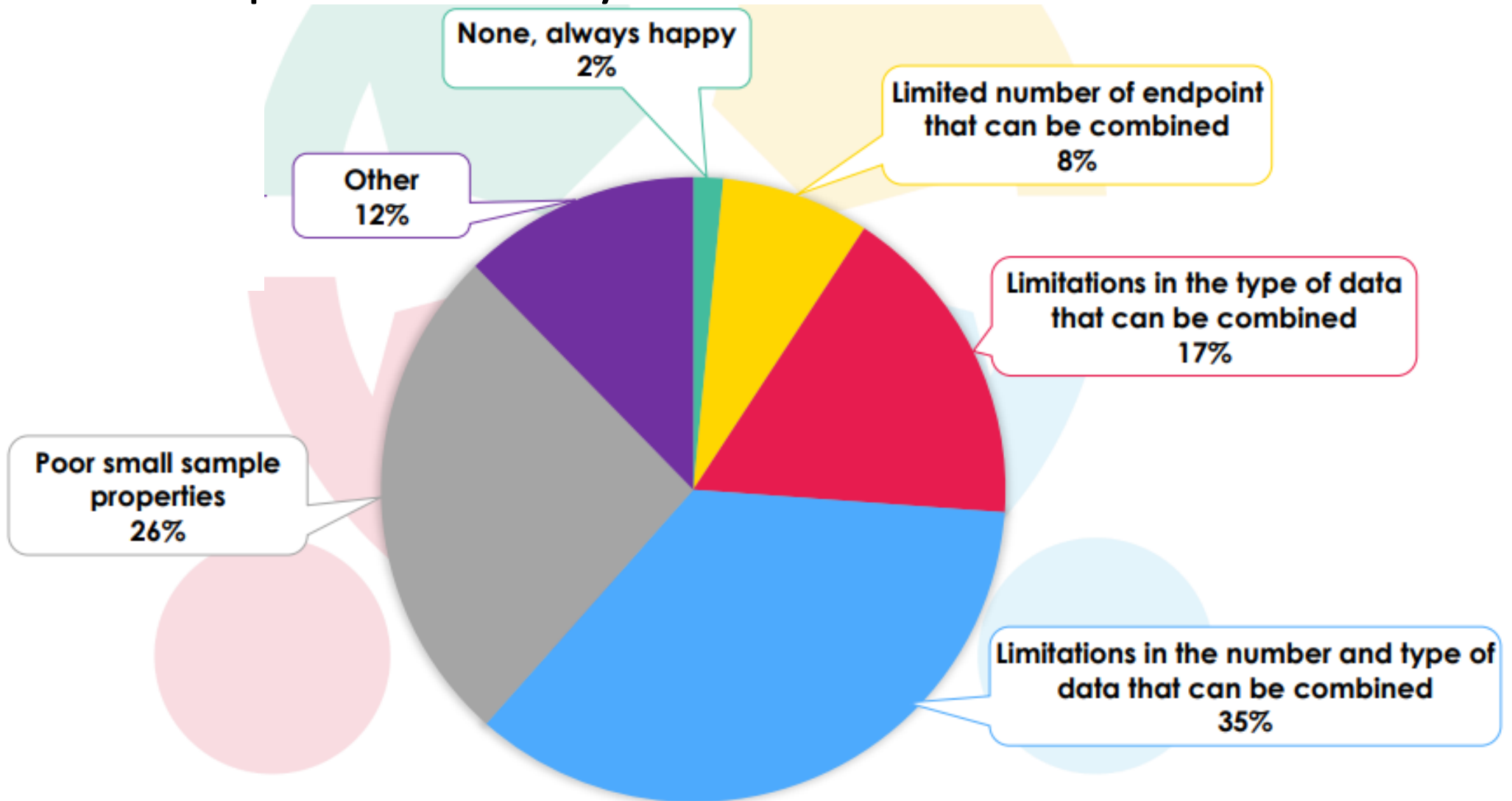
4. Dai et al. (2014)

5. Dmitrienko et al. (2010)

# Limitations of multivariate methods

- Ignore the correlation between the endpoints
- Limited to one type of endpoints
- Treats every endpoint as equally important
- No straightforward effect sizes measure to quantify the effect of the treatment is available
- Small sample properties

# What are the limitations you encounter with multiple endpoint analyses?



# Novel non-parametric methods

- **Based on ranks:**

Global rank<sup>1</sup>, Desirability of Outcome Ranking (DOOR)<sup>2</sup>; unambiguous ranks are not possible for multivariate censored outcomes

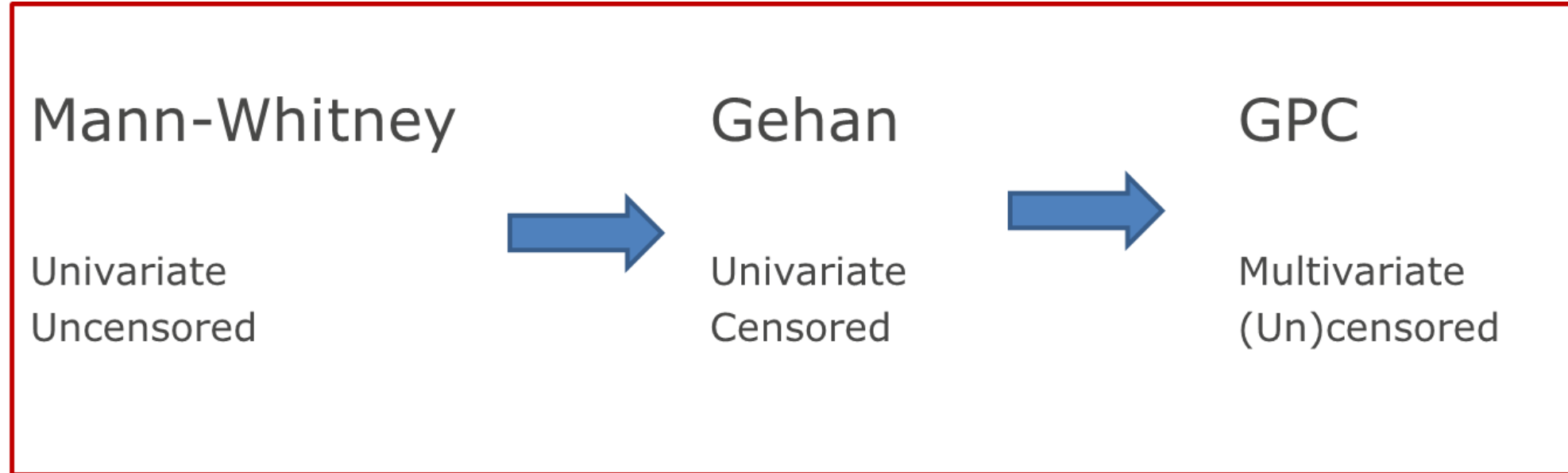
- **Extension of Mann-Whitney test :**

Generalized Pairwise Comparisons<sup>3</sup> (or win statistics<sup>4</sup>)

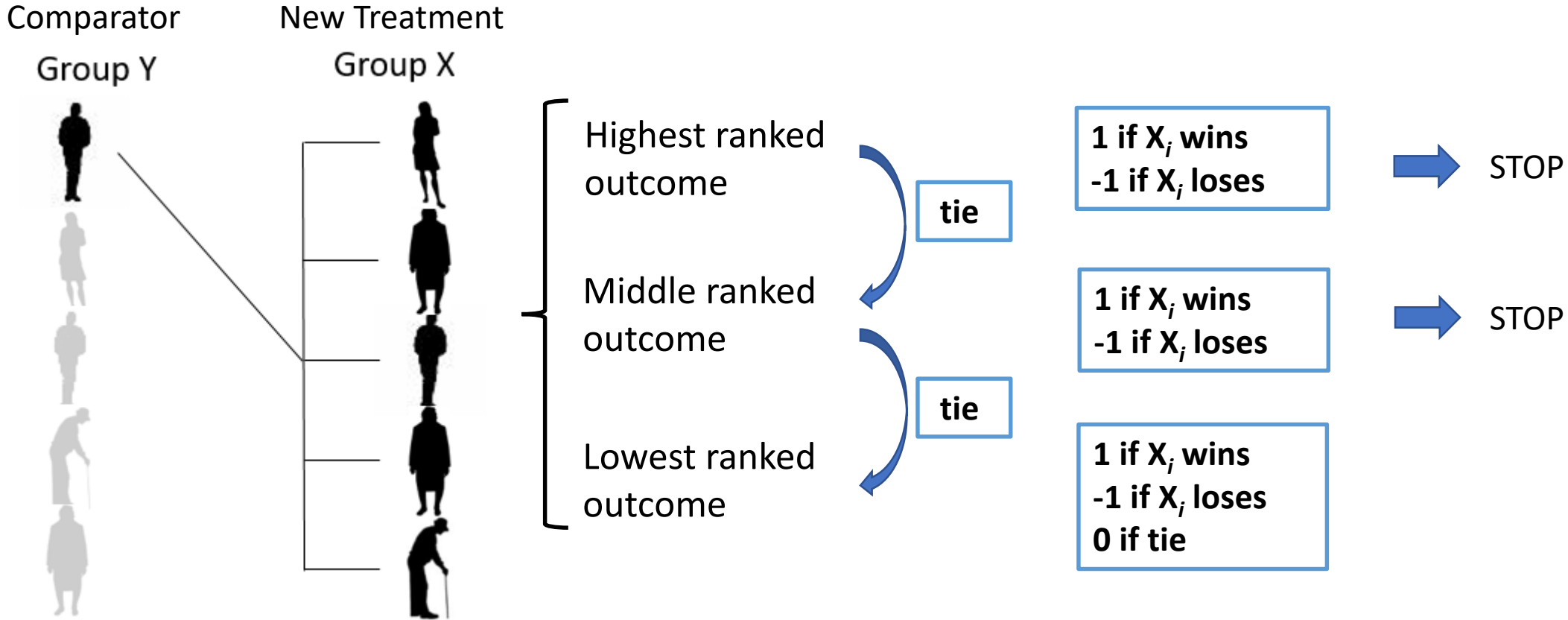
1. Felker and Maisel (2010)
2. Evans et al. (2015)
3. Buyse (2010)
4. Dong et al. (2021)

# Generalized Pairwise Comparisons (GPC)

# Family of GPC



# Generalized Pairwise Comparison (GPC) methodology



Finkelstein et al. (1999)  
Buyse (2010)  
Pocock et al (2012)

# GPC statistics

$$\text{Net (treatment) benefit} = \frac{N_X - N_Y}{nm} \leftarrow \text{Amount of pairs}$$

Number of wins for the treatment subjects

Number of wins for the control subjects

Net benefit ( $\Delta$ ): values between  $[-1, 1]$   
 $\Delta = P(X > Y) - P(X < Y)$

= U-statistic

Related to probabilistic index, relative effect, ... ( $\theta$ ):  
 $\theta = P(X > Y) + 1/2 P(X = Y)$

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

Buyse (2010)



# GPC statistics

$$\text{Net (treatment) benefit} = \frac{N_X - N_Y}{nm}$$

← Amount of pairs

Number of wins for the treatment subjects

Number of wins for the control subjects

$$\text{Win Ratio: } \frac{N_X}{N_Y}$$

Values between  $[0, \infty [$

Ignores ties



# GPC statistics

$$\text{Net (treatment) benefit} = \frac{N_X - N_Y}{nm} \leftarrow \text{Amount of pairs}$$

Number of wins for the treatment subjects

Number of wins for the control subjects

$$\text{Win Odds Ratio: } \frac{N_X + 1/2 N_{X=Y}}{N_Y + 1/2 N_{X=Y}} \leftarrow \text{Number of ties}$$

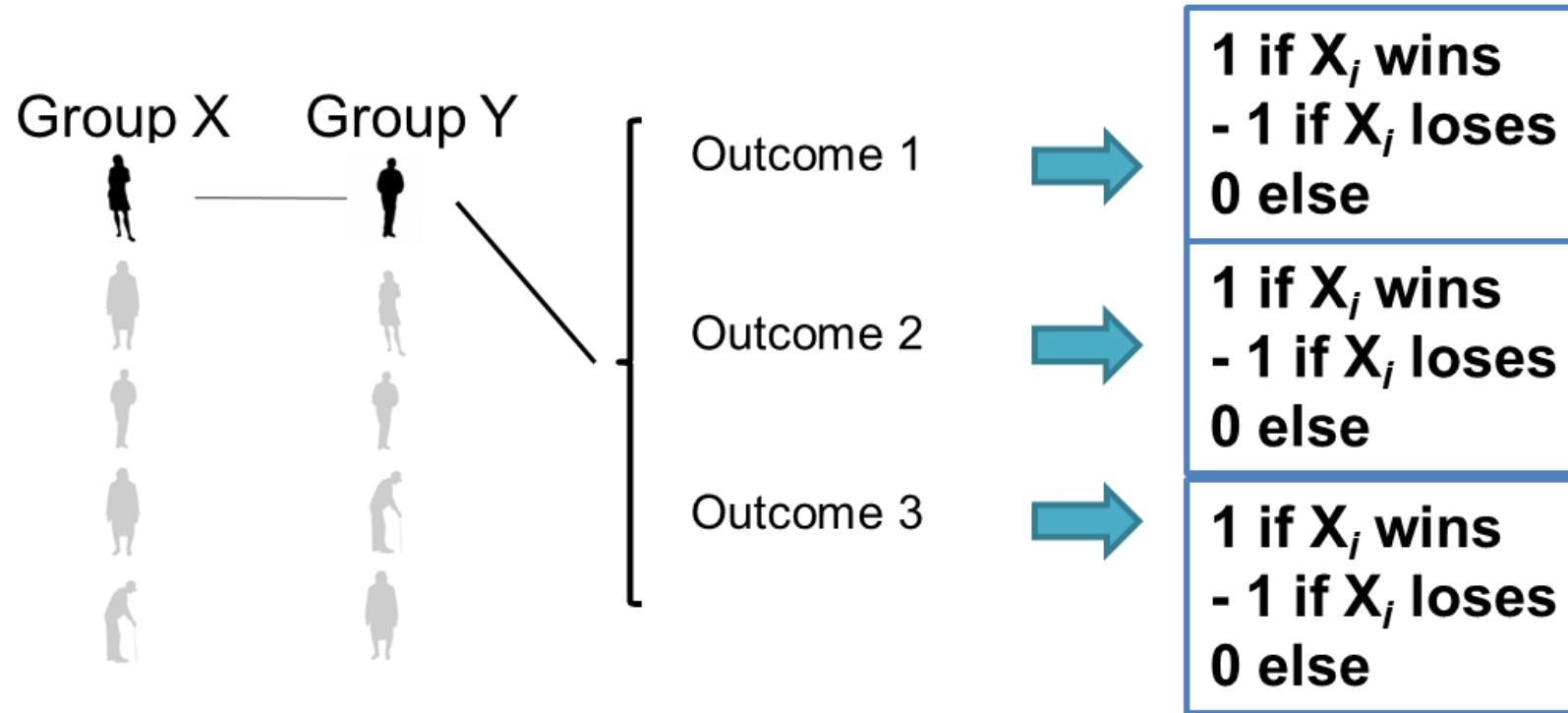
$$\text{Win Ratio: } \frac{N_X}{N_Y}$$

Values between  $[0, \infty [$

$$= \frac{1+\Delta}{1-\Delta}$$

Dong et al. (2020)  
Brunner et al. (2021)

# Non-prioritized GPC



$$\Delta = \frac{N_X - N_Y}{nmk}$$

with  $k$  the number of outcomes

O'Brien (1984)  
Ramchandani et al. (2016)  
Verbeeck et al. (2019)

# Flexible framework of GPC

- Prioritized/non-prioritized<sup>1,2</sup>
- Matched/unmatched pairwise comparisons<sup>3</sup>
- Threshold of clinical relevance ( $\tau$ )<sup>4</sup>

# Characteristics of GPC

- **Univariate uncensored:** unbiased and efficient in clinical trials scenarios<sup>1</sup>
- **Univariate censored:** drop-out bias can be corrected<sup>2</sup>
- **Multivariate:** correlation between outcomes affects prioritized and non-prioritized GPC differently<sup>3</sup>

# Inference with GPC

Net benefit	Win ratio	Win odds
Re-sampling permutation test	Re-sampling bootstrap test	Rank-based test
Asymptotic Normal U-statistic	Asymptotic Lognormal U-statistic	

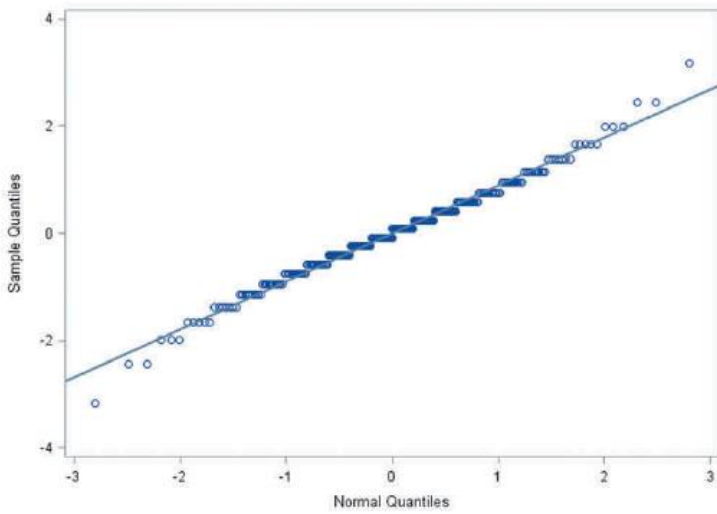
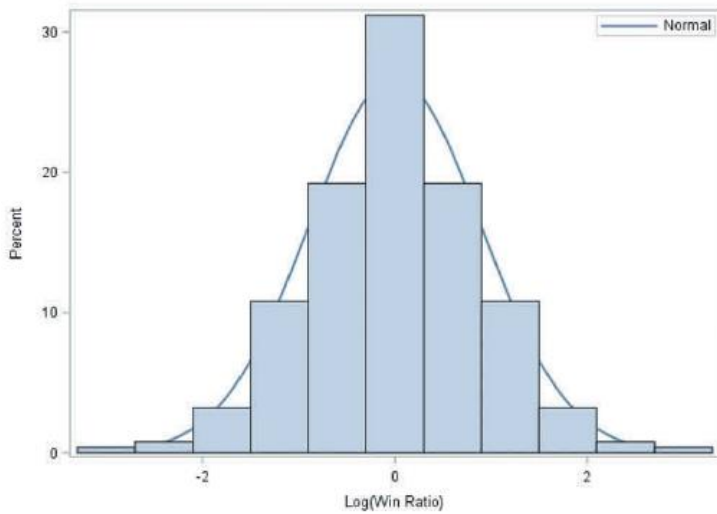
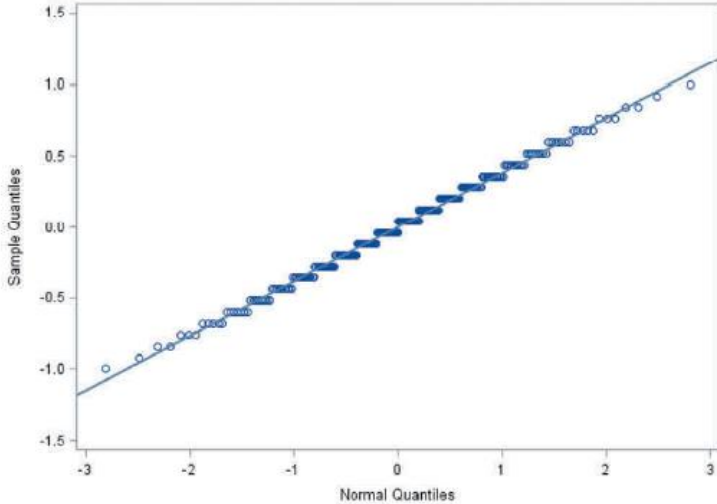
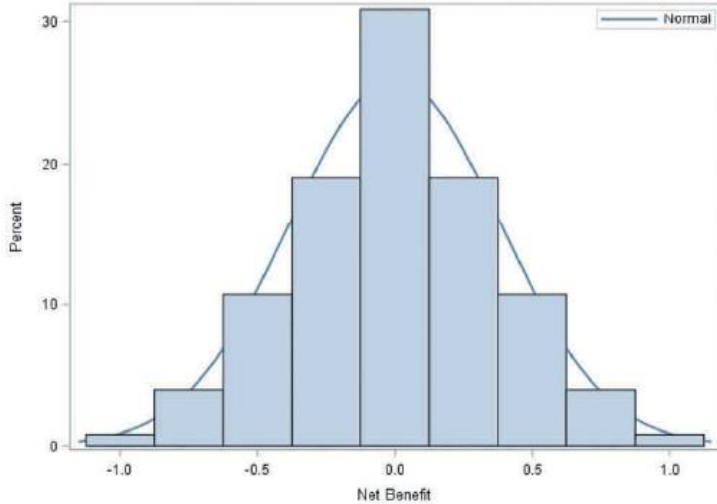
Theoretically shown that GPC test with net benefit, win ratio and win odds are approximately equal

## Small sample behavior?

# Small sample inference with GPC

- Extend **exact permutation test** of Gehan and Gilbert to win ratio, to bootstrap test and non-prioritized GPC.
- The null distribution of the GPC statistic in every possible permutation (bootstrap) sample is standard normally distributed.

# Small sample inference with GPC



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





# Small sample inference with GPC

Type I error

N	U-Statistic Ramchandani	U-Statistic Dong	U-statistic Bebu	Exact Permutation	Exact Bootstrap
20	0.0210	0.0099	0.1175	0.0512	0.0792
50	0.0390	0.0347	0.0717	0.0483	0.0599
100	0.0457	0.0436	0.0603	0.0507	0.0556
200	0.0486	0.0477	0.0548	0.0502	0.0522



# GPC corrects all limitations of multivariate methods

- Captures correlation between the endpoints
- Allows any number and type of endpoints
- Allows priority ranking of endpoints by severity
- Straightforward effect sizes measure to quantify the effect of the treatment
- Good small sample properties

# GPC method accepted by regulatory authorities

- Amyloid cardiomyopathy (ATTR-CM)
- Prevalence <1/100,000 in EU
- Accumulation of misfolded transthyretin amyloid fibrils in the myocardium, leading to restrictive cardiomyopathy and heart failure.
- Drug approval Vyndaqel (tafamidis) by FDA (May 2019) and EMA (Feb 2020) based on ATTR-ACT trial:
  - 441 patients
  - **Primary endpoint: GPC** with all-cause mortality, followed by cardiovascular-related hospitalizations

# Example: Epidermolysis bullosa

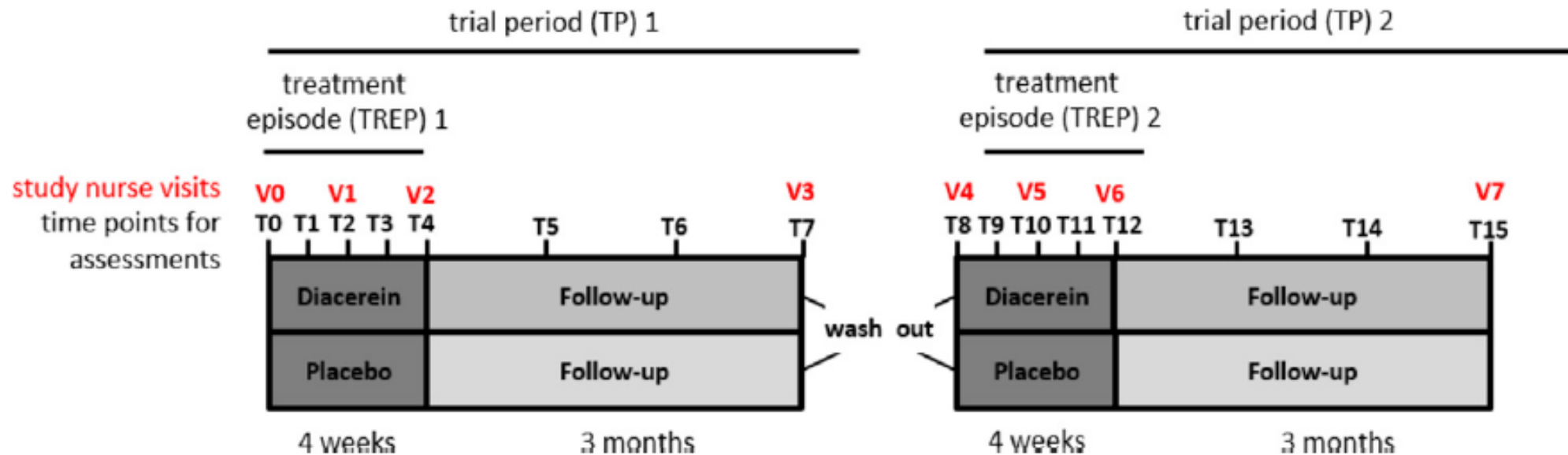
European Joint Programme on Rare Diseases:

“Demonstration projects on existing statistical methodologies to improve RD clinical trials”

EBStatMax project (Salzburg, Hasselt, Uppsala)

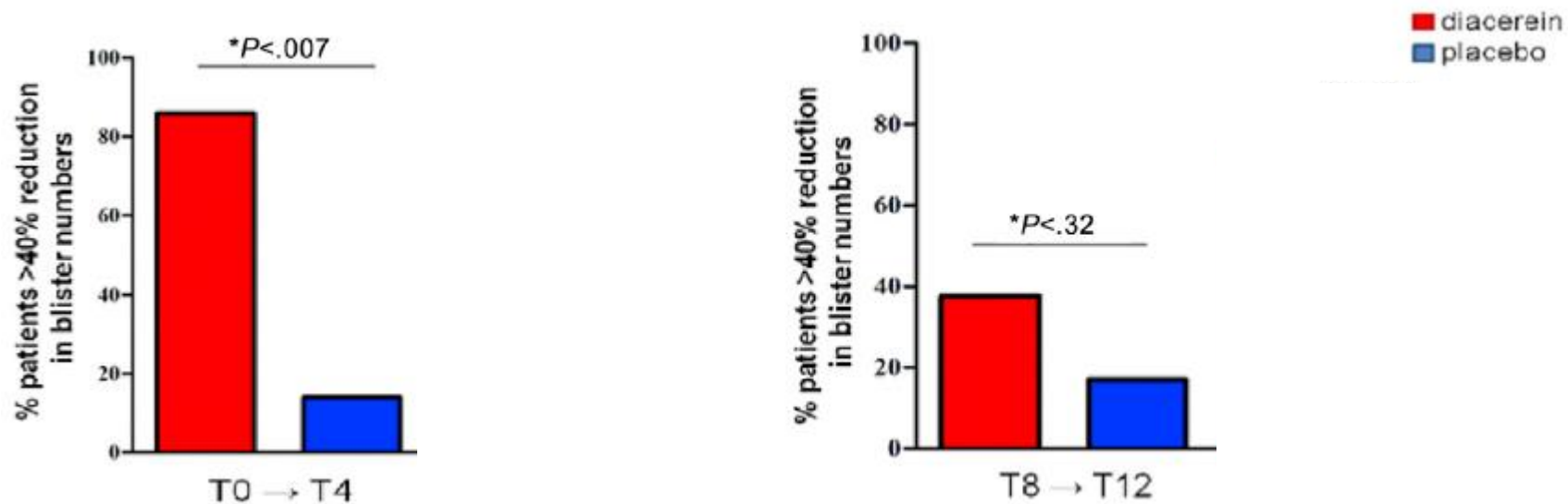
# EB trial design

- Rare skin disease: Epidermolysis bullosa simplex
- Formation of blisters under low mechanical stress
- 15 pediatric subjects (with missing data) treated with placebo and diacerein cream in a longitudinal cross-over trial



# Inconclusive results primary endpoint analysis

- Primary endpoint: >40% reduction in blister count compared to baseline (binary outcome) at week 4; Barnard test (~Fisher exact test 2x2 table)



# But.....

- Barnard test ignores:
  - Cross-over design
  - Longitudinal data: blister count measurement: 2, 4, weeks and 3 months
  - Patient relevant outcomes: QoL: baseline and post-treatment visit 4 weeks
  
- Question:

*“Is there a powerful test, accounting for the cross-over design and longitudinal information?”*

# Wide array of tests are being evaluated for blister outcome

- Non-parametric:
  - Rank-based marginal model for longitudinal data (nparLD)
  - GPC
- Semi-parametric
  - GEE-type model with small sample corrections
- Parametric
  - Model averaging



# GEE-type model performs best for cross-over longitudinal measures

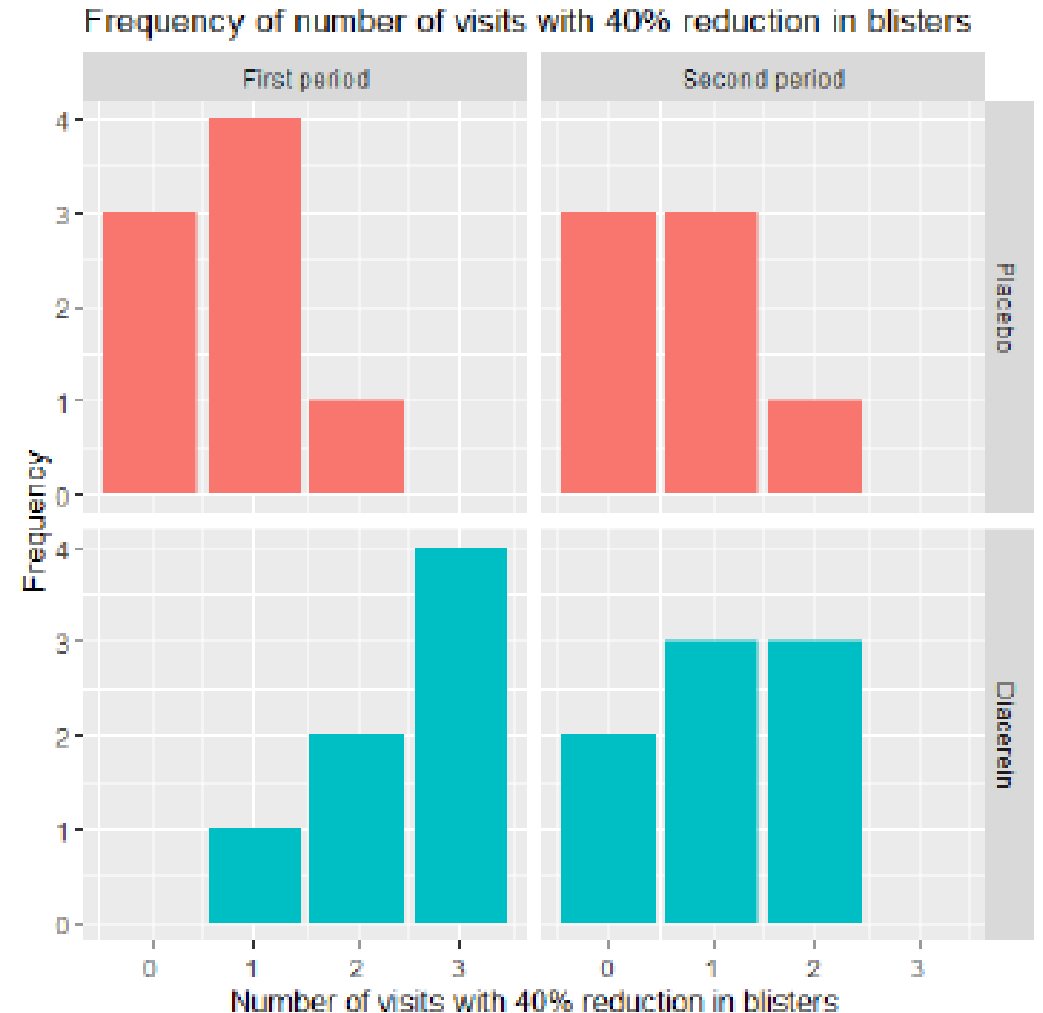
	One-sided			Two-sided		
	Samples	Type I	Power	Samples	Type I	Power
Barnard period 1	5000/5000	0.035	0.34	3971/4620	0.029	0.17
Barnard period 2	5000/5000	0.032	0.25	4675/4995	0.041	0.08
Marginal model period 1				4882/4966	0.069	0.15
Marginal model period 2				4999/4997	0.066	0.14
Matched univariate GPC	5000/5000	0.055	0.13	5000/5000	0.047	0.06
Unmatched univariate GPC	5000/5000	0.058	0.18	5000/5000	0.052	0.11
Matched prioritized GPC	5000/5000	0.016	0.05	5000/5000	0.077	0.03
Unmatched prioritized GPC	5000/5000	0.055	0.18	5000/5000	0.044	0.10
Unmatched non-prioritized GPC	5000/5000	0.058	0.21	5000/5000	0.059	0.13
GEE - no correction				4878	0.083	/
GEE - Kauermann & Carroll				4878/4679	0.071	0.54
GEE - Fay & Graubard				4878/4679	0.069	0.54
GEE - Mancl & DeRouen				4878/4679	0.054	0.55

$$\text{logit}(\pi_{ist}) = \beta_0 + \beta_1 G_{is} + \beta_2 P_i + \sum \beta_j T_{ist},$$

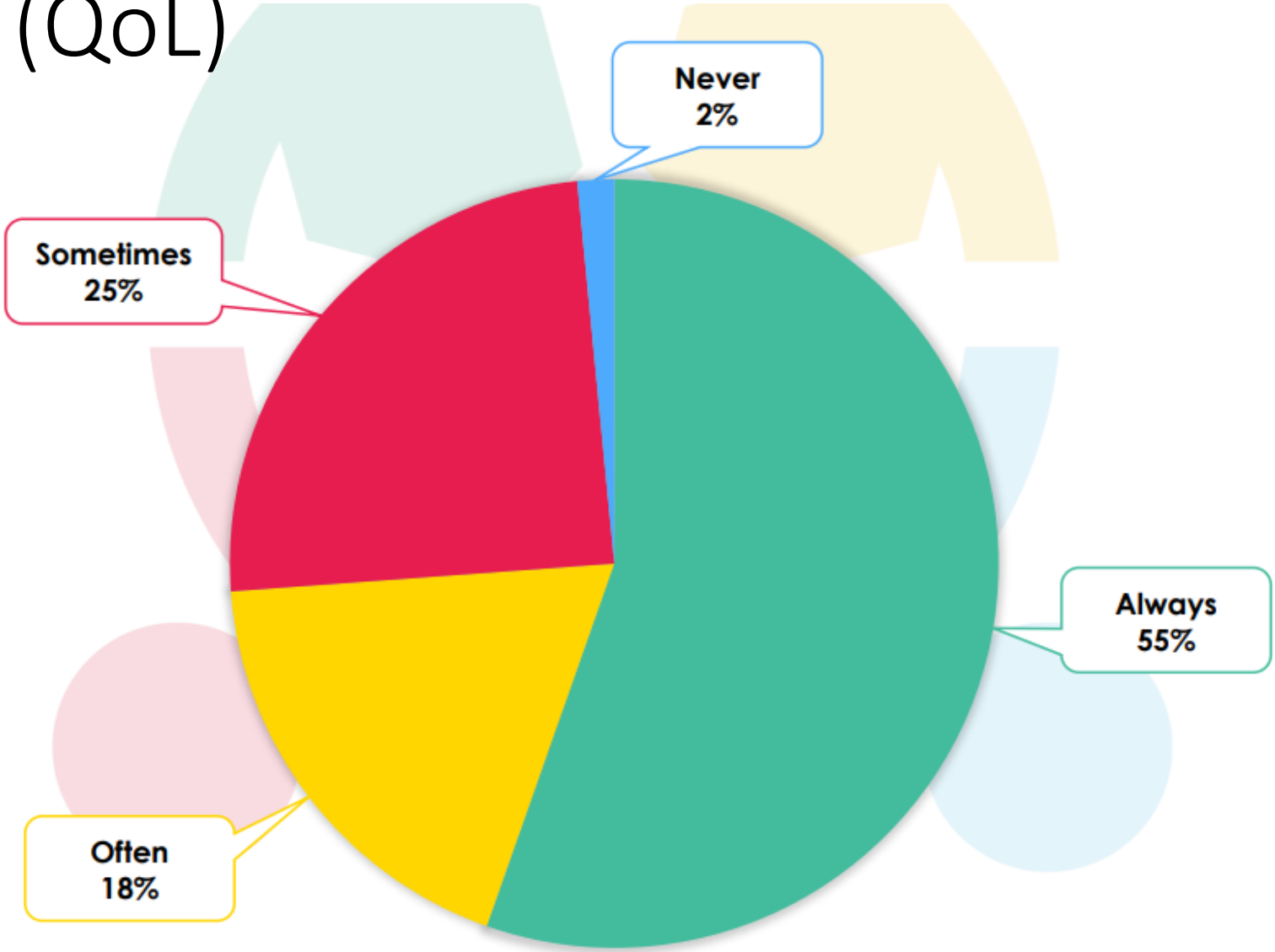
# Diacerin improves blister outcome

The **odds ratio** of a 40% reduction in the number of blisters between diacerein and placebo is **5.73 (95%CI: 1.50–21.91; *p-value* = 0.0125)**, which is mainly due to the effect in the first period.

The odds ratio of a 40% reduction in the number of blisters in period 1 versus period 2 is 4.34 (95% CI: 1.12–16.84; *p-value* = 0.0350).



# Multivariate outcome with patient reported outcome (QoL)

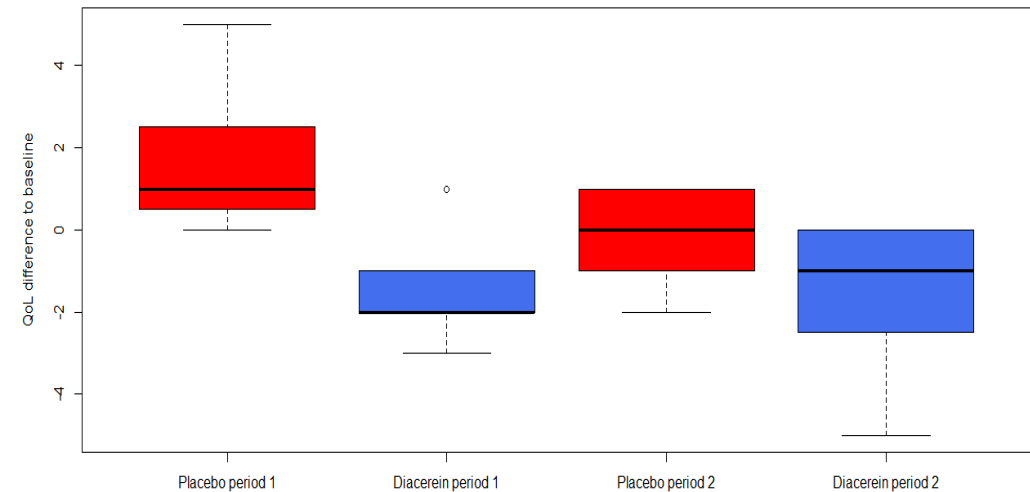
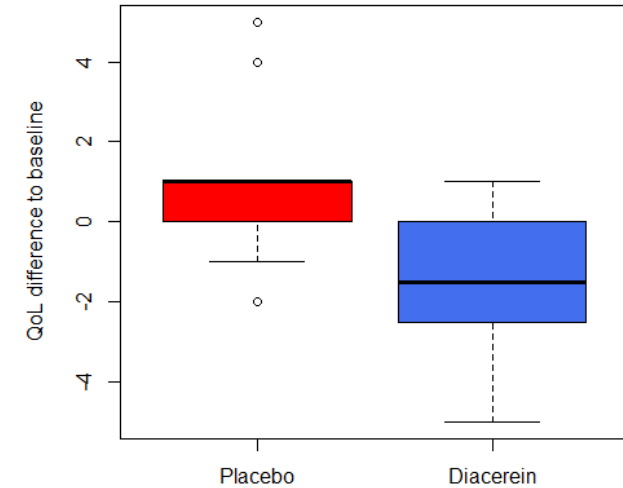


# Multivariate outcome with patient reported outcome (QoL)

QoL questionnaire on hindrance daily activities:

- 8 questions
- Each scored:  
0 (no hindrance)-3 (very much hindrance)
- Maximum of 24 points

Since QoL is measured only at baseline and post-treatment visit, we ignore the longitudinal profile of the blister outcome



# Multivariate outcome with patient reported outcome (QoL)

- Non-parametric:
  - Rank-based marginal model for longitudinal data (nparLD)
  - GPC
- Semi-parametric
  - GEE-type model with small sample corrections
- Parametric
  - Model averaging

# Variants of GPC

- (Unmatched) Prioritized GPC:
  - 40% blister reduction
  - QoL difference to baseline
- (Unmatched) Non-prioritized GPC
- Matched prioritized GPC

# Matched GPC inference

- Conditional sign test:

$$Z_m = \frac{N_X - N_Y}{\sqrt{N_X + N_Y}} \sim N(0,1)$$

Uniformly most powerful test

- But:
  - requires at least 15-20 (paired) subjects
  - ignores number of ties
  - Konietzschke and Pauly (2012) motivate that under certain conditions (applicable for the exact permutation test) the paired design can be ignored.

# Simulation set-up

- Permute EB trial blister count and QoL over treatment arms 5000 times
- Add a random Poisson( $\lambda=3$ ) treatment effect for both the placebo blister count and QoL for the placebo arm
- Dichotomized blister count (40% reduction) and standardized difference with baseline  $(\frac{y_0 - y_4}{y_0})$



# Matched GPC: often uncontrolled type I error

		Type I error	Power
		Dichotomized blister outcome	
N=13	unmatched blister	0.0692 (0.0216)	0.5904 (0.7202)
	unmatched QoL	0.0514 (0.0486)	0.8642 (0.9302)
	unmatched prioritized	0.0514 (0.0510)	0.9594 (0.9812)
	unmatched non-prioritized	0.0490 (0.0524)	0.9886 (0.9716)
	matched blister	0.0348 (0.0544)	0.4751 (0.0002)
	matched QoL	0.0422 (0.0550)	0.7044 (0.0000)
	matched prioritized	0.0260 (0.0258)	0.5824 (0.8210)
		Standardized difference blister outcome	
N=12	unmatched blister	0.0438 (0.0450)	0.5138 (0.6650)
	unmatched QoL	0.0490 (0.0528)	0.7940 (0.8888)
	unmatched prioritized	0.0442 (0.0458)	0.5402 (0.6852)
	unmatched non-prioritized	0.0510 (0.0502)	0.9250 (0.9670)
	matched blister	0.0472 (0.0654)	0.2784 (0.0004)
	matched prioritized	0.0414 (0.0724)	0.2714 (0.5440)

Two-sided (one-sided) type I error and power

# Adding QoL to blister increases power,...

	Type I error	Power
	Dichotomized blister outcome	
unmatched blister	0.0692 (0.0216)	0.5904 (0.7202)
unmatched QoL	0.0514 (0.0486)	0.8642 (0.9302)
unmatched prioritized	0.0514 (0.0510)	0.9594 (0.9812)
unmatched non-prioritized	0.0490 (0.0524)	0.9886 (0.9716)
matched blister	0.0348 (0.0544)	0.4751 (0.0002)
matched QoL	0.0422 (0.0550)	0.7044 (0.0000)
matched prioritized	0.0260 (0.0258)	0.5824 (0.8210)
	Standardized difference blister outcome	
unmatched blister	0.0438 (0.0450)	0.5138 (0.6650)
unmatched QoL	0.0490 (0.0528)	0.7940 (0.8888)
unmatched prioritized	0.0442 (0.0458)	0.5402 (0.6852)
unmatched non-prioritized	0.0510 (0.0502)	0.9250 (0.9670)
matched blister	0.0472 (0.0654)	0.2784 (0.0004)
matched QoL	0.0414 (0.0524)	0.6536 (0.0000)
matched prioritized	0.0414 (0.0724)	0.2714 (0.5440)

N=15x15

N=14x14

Two-sided (one-sided) type I error and power

... but less so for the prioritized continuous outcome

	Type I error	Power
	Dichotomized blister outcome	
unmatched blister	0.0692 (0.0216)	0.5904 (0.7202)
unmatched QoL	0.0514 (0.0486)	0.8642 (0.9302)
unmatched prioritized	0.0514 (0.0510)	0.9594 (0.9812)
unmatched non-prioritized	0.0490 (0.0524)	0.9886 (0.9716)
matched blister	0.0348 (0.0544)	0.4751 (0.0002)
matched QoL	0.0422 (0.0550)	0.7044 (0.0000)
matched prioritized	0.0260 (0.0258)	0.5824 (0.8210)
	Standardized difference blister outcome	
unmatched blister	0.0438 (0.0450)	0.5138 (0.6650)
unmatched QoL	0.0490 (0.0528)	0.7940 (0.8888)
unmatched prioritized	0.0442 (0.0458)	0.5402 (0.6852)
unmatched non-prioritized	0.0510 (0.0502)	0.9250 (0.9670)
matched blister	0.0472 (0.0654)	0.2784 (0.0004)
matched QoL	0.0414 (0.0524)	0.6536 (0.0000)
matched prioritized	0.0414 (0.0724)	0.2714 (0.5440)

Two-sided (one-sided) type I error and power

# Univariately: little evidence of a treatment effect

		# wins	#losses	#ties	Net Benefit (95%CI)	p-value one-sided	p-value two-sided
<b>Dichotomized blister outcome + QoL</b>							
matched univariate GPC QoL		9	0	4	0.6923(NA;NA)	0.0013	0.0027
matched prior GPC	Binary	5	2	6	0.2308 (-0.1716;0.5548)	0.1284	0.2568
	QoL	5	0		0.2		
	Overall	10	2	1	0.6154 (0.0879;0.8784)	0.0105	0.0209
unmatched prior GPC	Binary	99	24			0.3333	
	QoL	72	14		0.2578		
	Overall	171	38	16	0.5911 (0.1771;1.0000)	0.0026	0.0051
unmatched non-prior GPC	Binary	99	24	102		0.3333	0.0351
	QoL	162	22	41		0.6222	0.0010
	Overall				0.4778 (0.1719;0.7836)	0.0011	0.0022
<b>Standardized difference blister outcome + QoL</b>							
matched univariate GPC QoL		8	0	4	0.6667 (NA;NA)	0.0023	0.0047
matched prior GPC	Count	5	5	2	0 (-0.4525;0.4525)	0.5000	1.0000
	QoL	2	0		0.2		
	Overall	7	5	0	0.1667 (-0.3623;0.6148)	0.2819	0.5637
unmatched prior GPC	Count	130	61			0.3520	
	QoL	4	0		0.0204		
	Overall	134	61	1	0.3724 (-0.0628;0.8077)	0.0467	0.0935
unmatched non-prior GPC	Count	130	61	5		0.3520	0.0562
	QoL	141	19	36		0.6224	0.0013
	Overall				0.4872 (0.1482;0.8263)	0.0024	0.0049

# Multivariately: evidence of a treatment effect,...

		# wins	#losses	#ties	Net Benefit (95%CI)	p-value one-sided	p-value two-sided
<b>Dichotomized blister outcome + QoL</b>							
matched univariate GPC QoL		9	0	4	0.6923(NA;NA)	0.0013	0.0027
matched prior GPC							
	Binary	5	2	6	0.2308 (-0.1716;0.5548)	0.1284	0.2568
	QoL	5	0	0	0.2		
	<b>Overall</b>	<b>10</b>	<b>2</b>	<b>1</b>	<b>0.6154 (0.0879;0.8784)</b>	<b>0.0105</b>	<b>0.0209</b>
unmatched prior GPC							
	Binary	99	24			0.3333	
	QoL	72	14		0.2578		
	<b>Overall</b>	<b>171</b>	<b>38</b>	<b>16</b>	<b>0.5911 (0.1771;1.0000)</b>	<b>0.0026</b>	<b>0.0051</b>
unmatched non-prior GPC							
	Binary	99	24	102		0.3333	0.0351
	QoL	162	22	41		0.6222	0.0010
	<b>Overall</b>				<b>0.4778 (0.1719;0.7836)</b>	<b>0.0011</b>	<b>0.0022</b>
<b>Standardized difference blister outcome + QoL</b>							
matched univariate GPC QoL		8	0	4	0.6667 (NA;NA)	0.0023	0.0047
matched prior GPC							
	Count	5	5	2	0 (-0.4525;0.4525)	0.5000	1.0000
	QoL	2	0	0	0.2		
	<b>Overall</b>	<b>7</b>	<b>5</b>	<b>0</b>	<b>0.1667 (-0.3623;0.6148)</b>	<b>0.2819</b>	<b>0.5637</b>
unmatched prior GPC							
	Count	130	61			0.3520	
	QoL	4	0		0.0204		
	<b>Overall</b>	<b>134</b>	<b>61</b>	<b>1</b>	<b>0.3724 (-0.0628;0.8077)</b>	<b>0.0467</b>	<b>0.0935</b>
unmatched non-prior GPC							
	Count	130	61	5		0.3520	0.0562
	QoL	141	19	36		0.6224	0.0013
	<b>Overall</b>				<b>0.4872 (0.1482;0.8263)</b>	<b>0.0024</b>	<b>0.0049</b>

... mainly in first treatment period

		# wins	#losses	#ties	Net Benefit (95%CI)	p-value one-sided	p-value two-sided
<b>Period 1</b>							
unmatched prior GPC							
	Bin	30	3		0.4821		
	QoL	17	0		0.3036		
	<b>Overall</b>	<b>47</b>	<b>3</b>	<b>6</b>	<b>0.7857 (0.2079;1.3635)</b>	<b>0.0038</b>	<b>0.0077</b>
unmatched non-prior GPC							
	Bin	30	3	23	0.4821	0.0331	0.0662
	QoL	43	2	11	0.7321	0.0038	0.0076
	<b>Overall</b>				<b>0.6071(0.1261;1.0882)</b>	<b>0.0067</b>	<b>0.0134</b>
<b>Period 2</b>							
unmatched prior GPC							
	Bin	18	5		0.2321		
	QoL	16	9		0.125		
	<b>Overall</b>	<b>34</b>	<b>14</b>	<b>8</b>	<b>0.3571 (-0.2346;0.9489)</b>	<b>0.1184</b>	<b>0.2368</b>
unmatched non-prior GPC							
	Bin	18	5	33	0.2321	0.1636	0.3271
	QoL	16	9	31	0.4464	0.0693	0.1385
	<b>Overall</b>				<b>0.3393(-0.0737;0.7522)</b>	<b>0.0537</b>	<b>0.1073</b>

# Conclusions

# Conclusions

- The GPC methodology is very flexible.
- It allows for a combination of any type and any number of outcomes, including patient relevant outcomes.
- Takes account of the correlation between outcomes.
- May increase power, compared to a univariate outcome.
- Allows for an easy interpretable treatment effect and gives insight into the partial contribution of outcomes to the overall result.
- The exact permutation is easy, fast and precise even in very small samples (Available in SAS, R and under development in Python).



# Questions ?



Interuniversity Institute for Biostatistics  
and statistical Bioinformatics



Johan Verbeeck

[johan.verbeeck@uhasselt.be](mailto:johan.verbeeck@uhasselt.be)

Data Science Institute/I-Biostat

UHasselt - Belgium

# References

- Barnard, G.A. (1947). Significance tests for 2×2 tables. *Biometrika*, 34:123–138.
- Buyse, M. (2010). Generalized pairwise comparisons of prioritized outcomes in the two sample problem. *Statistics in Medicine*, 29:3245–3257.
- Coakley, C.W., et al. (1996). Versions of the sign test in the presence of ties. *Biometrics* 52, 1242-1251.
- Fagerland, M., et al. (2013). The McNemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional. *BMC Medical Research Methodology*, 13:91.
- Konietzschke, F., et al. (2012). A studentized permutation test for the nonparametric Behrens-Fisher problem in paired data. *Electronic Journal of Statistics*. 6:1358–1372.
- O'Brien, P. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics*, 40(4):1079–1087.

# References

- Pocock et al. (2012). The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. *European Heart Journal*, 33:176–182.
- Verbeeck, J., et al. (2019) Generalized pairwise comparison methods to analyze (non)prioritized composite endpoints. *Statistics in Medicine*, 38:5641– 5656.
- Verbeeck, J., et al. (2020) Evaluation of inferential methods for the net benefit and win ratio statistics. *Journal of Biopharmaceutical Statistics*, 30(5):765-782.
- Verbeeck, J., et al. (2021) Unbiasedness and efficiency of non-parametric and UMVUE estimators of the probabilistic index and related statistics. *Statistical Methods in Medical Research*, 30(3), 747-768.
- Wally, V., et al. (2018). Diacerein orphan drug development for epidermolysis bullosa simplex: A phase 2/3 randomized, placebo-controlled, double-blind clinical trial. *J Am Acad Dermatol*. 78:892-901.